2021

Emotional body language synthesis for humanoid robots

Marmpena, Asimina

UNIVERSITY OF
PLYMOUTH

EMOTIONAL BODY LANGUAGE SYNTHESIS FOR HUMANOID ROBOTS

by

ASIMINA MARMPENA

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering, Computing and Mathematics

May 2021

'It has to be half laughing, half serious; with great splashes of exaggeration.'

— Virginia Woolf, December 1927

Acknowledgements

I would like to thank my supervisor, Angelica Lim, who has been there for me from the beginning of the PhD. Her advice, encouragement and endless support have been precious. She introduced me to the fascinating world of affective robotics, and she has been a real inspiration to me. I would also like to express my gratitude and appreciation to Nicolas Hemion, who listened to my ideas with attention and inspired me to work with the VAE framework. He facilitated my work in many ways, with his hands-on expertise.

Many thanks to Torbjørn Dahl for his valuable advice, and the fact that he enabled me to follow my ideas and did his best to support me. I am also grateful to Thomas Wennekers; his support, encouragement and advice were invaluable. I would also like to thank Angelo Cangelosi and Tony Belpaeme for their valuable feedback and their support.

Thanks to the APRIL project fellows for sharing tough times, joys, and lots of fun: Oksana, Bahar, Alex, and Pontus. I'm grateful for the amazing friends and colleagues with whom I spent the first half of my PhD in Plymouth: Sarah, Martina, François, Barış, Marta, Daniel, Riccardo, Frederico, Clara, Leszek, Debora, Ricardo, Emmanuel, Ilaria and Frank.

I am also grateful to many people at SoftBank Robotics, Paris, where I spent the rest of my PhD time. Thanks to Alban Laflaquière for his valuable insights and Alex Mazel for being incredibly supportive. A huge thanks to my good friend and collaborator Ferran Garcia, with whom I conducted the final user study. I'm also grateful to other SBRE and AI Lab people who have always been helpful and inspiring: Marc Moreaux (beloved master of lab fun during deadline crises), Mehdi A.B. (for great advice, inspiration and intelligent discussions), Giuseppe, Luis, Marwa, Luca, Mehdi H., Hugo, Michael. A big thanks to Natalia Lyubova, who made my first days in Paris less stressful with her fantastic support.

Finally, I would like to express my acknowledgements to friends and family in Athens, where I returned to write this thesis. Thanks to my lovely friends Elena, Eleni, Maria, Vaso, and Tassos for being supportive and always there for me when I needed to talk or enjoy a break. A big thanks to Margarita, Vasilis and Nadia for being next to me during this endeavour, supporting me in every possible way. Mostly thanks to Rita, who could always sense and instantly eliminate the doubts and fears of a first-gen doctorate student.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee. Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

This work has been carried out by Asimina Marmpena under the supervision of Dr Thomas Wennekers, Dr. Torbjørn S. Dahl, Dr. Angelica Lim, Dr. Nikolas Hemion, Dr. Alban Laflaquière, and Prof. Angelo Cangelosi. The work was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 674868 (APRIL) and carried out in collaboration with SoftBank Robotics, Europe, as an industrial partner.

Publications:

Marmpena M., Lim, A., and Dahl, T. S. (2018). *How does the robot feel? Perception of valence and arousal in emotional body language.* Paladyn, Journal of Behavioral Robotics, 9(1), 168-182. DOI: https://doi.org/10.1515/pjbr-2018-0012

Marmpena M., Lim, A., Dahl, T. S., and Hemion, N. (2019). *Generating robotic emotional body language with Variational Autoencoders.* In Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 545–551. DOI:10.1109/ACII.2019.8925459

Marmpena M., Garcia, F., and Lim, A. (2020). *Generating robotic emotional body language of targeted valence and arousal with Conditional Variational Autoencoders.* In Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, HRI '20, page 357–359. DOI: https://doi.org/10.1145/3371382.3378360

Word count for the main body of this thesis: **38,496**

Signed: Asimina Marmpena **Date:** 30 April 2020

Abstract

Emotional body language synthesis for humanoid robots
Asimina Marmpena

In the next decade, societies will witness a rise in service robots deployed in social environments, such as schools, homes, or shops, where they will operate as assistants, public relation agents, or companions. People are expected to willingly engage and collaborate with these robots to accomplish positive outcomes. To facilitate collaboration, robots need to comply with the behavioural and social norms used by humans in their daily interactions. One such behavioural norm is the expression of emotion through body language.

Previous work on emotional body language synthesis for humanoid robots has been mainly focused on hand-coded design methods, often employing features extracted from human body language. However, the hand-coded design is cumbersome and results in a limited number of expressions with low variability. This limitation can be at the expense of user engagement since the robotic behaviours will appear repetitive and predictable, especially in long-term interaction. Furthermore, design approaches strictly based on human emotional body language might not transfer effectively on robots because of their simpler morphology. Finally, most previous work is using six or fewer basic emotion categories in the design and the evaluation phase of emotional expressions. This approach might result in lossy compression of the granularity in emotion expression.

The current thesis presents a methodology for developing a complete framework of emotional body language generation for a humanoid robot, intending to address these three limitations. Our starting point is a small set of animations designed by professional animators with the robot morphology in mind. We conducted an initial user study to acquire reliable dimensional labels of valence and arousal for each animation. In the next step, we used the motion sequences from these animations to train a Variational Autoencoder, a deep learning model, to generate numerous new animations in an unsupervised setting. Finally, we extended the model to condition the generative process with valence and arousal attributes, and we conducted a user study to evaluate the interpretability of the animations in terms of valence, arousal, and dominance. The results indicate moderate to strong interpretability.

Table of contents

List of figures

List of tables

Nomenclature, notation and definitions

Acronyms / Abbreviations

AEVB	Autoencoding Variational Bayes
ANOVA	Analysis of Variance
CNN	Convolutional Neural Networks
CVAE	Conditional Variational Autoencoder
DAG	Directed Acyclic Graph
EBL	Emotional Body Language
ECA	Embodied Conversational Agent
ELBO	Evidence Lower Bound
GAN	Generative Adversarial Network
HCI	Human-Computer interaction
HRI	Human-Robot interaction
ICC	Intraclass Correlation
IRF	International Federation of Robotics
KDE	Kernel Density Estimates
LD	Latent Dimension
MCMC	Markov chain Monte Carlo
MLP	Multilayer Perceptron

Nomenclature, notation and definitions

MSE	Mean Square Error
PCA	Principal Component Analysis
SGD	Stochastic Gradient Descent
SGVB	Stochastic Gradient Variational Bayes
VAE	Variational Autoencoder

Notation

y	Variables in italic denote scalars. In particular, y is typically used to denote the scalar label of a training example
$\boldsymbol{x}, \boldsymbol{z}$	Variables in bold italic denote either multivariate random variables or a vector instance of them, e.g., a single observation (training example). In particular, $\boldsymbol{x}$ is used to denote instances of observed data, while $\boldsymbol{z}$ denotes latent variables (encodings of the observed data)
$\boldsymbol{x} \odot \boldsymbol{y}$	Element-wise multiplication of two vectors with equal size
θ, ϕ	The unknown parameters *theta* of the generative model (decoder) and *phi* of the inference model (encoder) respectively. The latter are also called variational parameters. Their values are learned via training
$p(\boldsymbol{x})$	Probability density functions (PDFs) or simply called *distributions*. We only refer to continuous distributions in this work
$p(\boldsymbol{x}, \boldsymbol{z})$	Joint probability distribution
$p(\boldsymbol{x} \mid \boldsymbol{z})$	Conditional probability distribution, i.e., the probability of $\boldsymbol{x}$ given $\boldsymbol{z}$

Definitions

$\mathbb{E}[g(x)]$	The expected value (expectation) of a function of a continuous random variable is $g\ \mathbb{E}[g(x)] = \int g(x)q(x)dx$
$D_{KL}(q \parallel p)$	The Kullback-Leibler divergence between two continuous distributions is $D_{KL}(q \parallel p) = \mathbb{E}_{x\sim q}[\log q(x) - \log p(x)]$

Bayes' rule	The posterior $p(a \mid b)$ is equal to the likelihood $p(b \mid a)$ multiplied by the prior $p(b)$ and normalized by the marginal distribution (evidence) $p(a)$, i.e., $p(a \mid b) = p(b \mid a)p(b)/p(a)$
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	A multivariate Normal distribution (Gaussian) with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

Chapter 1

Introduction

1.1 Social robotics

According to the International Federation of Robotics (IRF) report published in 2018[1], from 2019 through 2021 service robots market is anticipated to grow significantly. In particular, 1.67 million robots are expected to be sold for education and research, 10.7 million for entertainment, and 34,400 for elderly and handicap assistance. Regarding the last category, the report maintains that although the market share is currently small, a substantial increase can occur within the next 20 years, by an average of 29% per year to a value of around US$126 million in 2022. According to the same report, another rapidly growing category of service robots is public relation robots and robot assistants, used in hotels, museums, restaurants, shops, etc., to assist, guide, provide information, monitor guests, visitors, or clients. The sales of public relation robots could reach 93,350 units from 2019 to 2021.

Evidently, more and more service robots will get deployed in social environments like homes, schools, healthcare and commercial facilities. Consequently, their ability to interact with humans in a social way is becoming increasingly crucial [34]. Social robots must communicate with humans in an interpersonal manner, engage humans as partners, collaborate or coordinate with them to accomplish positive outcomes [35] in education, therapy and healthcare, retail trade, accommodation and food services, and so forth.

The preceding facts constitute compelling reasons and key drivers for the emergence and progress of a relatively new field of research, social robotics. According to Bartneck et al. [17]:

[1] https://ifr.org/post/market-for-professional-and-domestic-service-robots-booms-in-2018

> "A social robot is an autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioural norms expected by the people with whom the robot is intended to interact."

This definition excludes non-embodied agents such as screen characters and avatars. It requires some degree of autonomy, in the sense that if a human completely teleoperates a robot, then it is merely seen as an extension of a human and not as a social agent. Furthermore, it sets human-robot interaction (HRI) and communication as a precondition, excluding robots that only interact with other robots. Most importantly, for a robot to qualify as social, it must be equipped with the ability to perceive, understand and mimic aspects of human behaviour, as well as societal and cultural norms and values.

Fong et al. [73] identify the following characteristics for socially interactive robots:

- express and/or perceive emotions;
- communicate with high-level dialogue;
- learn/recognize models of other agents;
- establish/maintain social relationships;
- use natural cues (gaze, gestures, etc.);
- exhibit distinctive personality and character;
- may learn/develop social competencies.

It appears from this list of characteristics that the ability to sense emotional states and mimic emotional expressions is an essential feature for a socially interactive robot.

1.2 Affective robotics and computing

Affective robotics is an emerging subfield of social robotics. It mainly investigates how to enhance social robots with artificial emotional intelligence and how humans perceive such robots. It also examines ethical issues that emerge when such abilities become more and more sophisticated. Affective robotics largely overlaps with affective computing, a branch of computer science pioneered by Rosalind Picard [184] which extends the study of artificial emotional intelligence to non-embodied agents and machines. Affective computing is a scientific and engineering endeavour focusing on affect detection and generation. The field seeks to equip machines with the ability to recognise emotions from human affect signals

(e.g., facial expressions, body gestures, speech or neurophysiological signals) and adapt their behaviours appropriately according to such signals. In a few words, affective computing aims to build machines with artificial emotional intelligence that can simulate empathy.

According to Fong et al. [73], artificial emotions in social robotics are essential for several reasons. Firstly, they can render the robot as a believable character. In the context of social robotics, believability is related to the literary term "suspension of disbelief", i.e., to intentionally and rationally suspend judgement concerning the implausibility of an artefact. Believability is indispensable for engaging users in long-term interaction with artificial agents [18], and appropriate displays of robotic affect can drastically augment it [137]. Another use of artificial emotion expression is to indicate a robot's internal state, goals or intentions. For instance, MacLennan et al. [147] suggested that a robot could trigger the expression of fear as a feature of a self-protection mechanism. In a similar vein, Lim and Okuno proposed two gut feeling states, flourishing and distress, tied to a robot's physical state: flourishing corresponding to full battery and CPU/motor temperatures within working limits, and distress corresponding to a near-empty battery or hot motors [142]. A third use proposed by Fong et al. [73] employs emotions as a control mechanism, i.e., to determine precedence among different behaviours, planning, learning and adaptation in complex environments. Consequently, although emotion expression for a robot might appear as a redundant signal compared to a more fundamental response, it might be a crucial feature in particular use-cases. In naturalistic social HRI, robotic expression of emotion can render the robot as more believable and likeable and the interaction as more enjoyable [172, 136, 197] and engaging.

Emotion expression can also enhance human-computer interaction (HCI) with other artificial agents such as Embodied Conversational Agents (ECAs). ECAs [47] are animated screen characters, usually of human-like embodiment, designed to interact with human users through dialogue and nonverbal behaviours. Such agents' nonverbal behaviours can exploit various modalities, such as facial expression, eye gaze, speech intonation, and body language. Essentially, ECAs are computer interfaces that can improve user satisfaction and engagement since they enable a more naturalistic face-to-face interaction with a computer system [76]. ECAs can be deployed in several application scenarios, such as computer games, education, computer-assisted therapy and customer service. Social presence is by definition a core objective in designing compelling ECAs, and in accomplishing it, nonverbal affective communication is equally essential to verbal skills [47, 177, 77].

Arbib and Fellous [4] distinguish two aspects of emotion: the internal aspect, in which emotion facilitates the organization of behaviour (action selection, attention, and learning), and the external aspect, in which emotion acts as a means of communication and social integration. They note that in animals, these aspects are developed through a co-evolving

process. Although such a process is of great interest to affective robotics, another approach is working independently on either the internal or the external aspect of emotion, and potentially integrate the outcomes in a higher level of abstraction.

1.3 Emotional body language

The human cognition has evolved to frequently scan other humans to detect nonverbal signals that will update our expectations regarding other people's emotions, feelings, and moods. Such a process can be conscious, under the focus of attention, or operating in the background, perceived subconsciously. Combined with the tendency to attribute anthropomorphic traits to non-human agents [95, 57, 127], this innate human behaviour can explain why robots endowed with the ability to display nonverbal affective signals can be more appealing. Emotional expression can increase a robot's believability and improve engagement and collaboration in human-robot interaction [100].

Nonverbal affective signals in humans include facial expression, body gestures, deictic gestures, eye gaze direction, non-linguistic sounds, speech features, etc. Facial expression is by far the most widely studied nonverbal signal of affect in psychology literature, with affective bodily expression lagging far behind [116]. Nevertheless, bodily affect constitutes a powerful communication channel of great importance for conveying and perceiving emotion among humans [123, 229].

According to Fast [71], body movement and posture comprise the most primitive behavioural channel for humans to express emotion, and compared to facial expression, it can reveal more about the actual affective state [5, 42, 71, 111]. In terms of trustworthiness, a bodily expression can reflect a person's affect state more reliably, even when it is contradicted by facial expression and verbal communication [5]. Gestures and body postures evoke more trust [56], and they are less susceptible to social editing compared to facial expression, since people are more aware, and thus more in control, of their facial expression [67]. Furthermore, studies have shown that changes in a communicator's body orientation, posture configuration, proxemics, etc. can influence the overall likeability, interest, and openness between individuals [5, 42, 163, 111]. Bodily expression of emotion has also been found more effective than facial expression in discriminating between intense positive and negative emotions [6]. Finally, conveying emotion through the body has an additional practical advantage because the communicator's affect state is more easily visible from a distance [61, 221].

Consequently, it is anticipated that the integration of appropriate emotional body language (EBL) as a robotic behaviour during human-robot interaction can enhance the user experience and make it more engaging. Evidence from studies in HRI indicates that bodily expressions of

affect can draw human's attention [197] and they are essential in naturalistic social interaction for robots that lack expressive faces (appearance-constrained) [26]. Furthermore, emotion or personality expressiveness can make the interaction more enjoyable [136, 172]. Robots with the ability to mimic emotional expressions with their body language could be deployed in various human-robot interaction settings and applications, be it human-robot collaboation tasks, robot-assisted therapy with kids or elderly, education, entertainment, public relation, etc.

1.4 The affective loop framework

The affective loop [106] framework is a contextual abstraction initially proposed for the design of affective interaction systems and later adapted for emotion modelling in affective robotics by Paiva et al. [180]. In the initial framework, the affective loop is described as an experience between a user and a system which can influence and be influenced by the user. This process involves certain experiential qualities that arise through interaction, and it can be accomplished by building the system with the ability to mirror physical, cultural or social features of human embodied experiences while, at the same time, leaving enough room for users to interpret the system's response through their own lens of emotion, as active co-constructors of meaning [106].

In Fig. 1.1, we present the affective loop flow as proposed for robotic emotion modeling by Paiva et al. [180]. The robot detects human affect, possibly via facial expression, speech features, sentiment analysis, and other affect recognition channels. Subsequently, this information is used for generating the robot's emotional behaviour, which influences the user, elicits an updated state of human affect, and then a new cycle begins. Robot's emotional behaviour generation comprises three modules: emotion synthesis, emotion adaptation, and emotion expression. In our interpretation, these three successive modules can be implemented as independent modules, which only exchange information through some common data representation, or could overlap to some degree. For the emotion synthesis module, several emotional architectures have been proposed, some inspired by psychological or neurobiological models of emotion, and some completely data-driven ones (for references see [180, 192, 143, 142, 207, 167]). The emotion adaptation module is described as a mechanism for empathy. It takes into account user personalisation to adjust the output of the emotion synthesis module. Finally, the adapted or personalised synthetic emotion is passed to the emotion expression module to display the appropriate robotic behaviour.

Paiva et al. [180] distinguish three important purposes or roles that must be served in affective human-robot interaction:

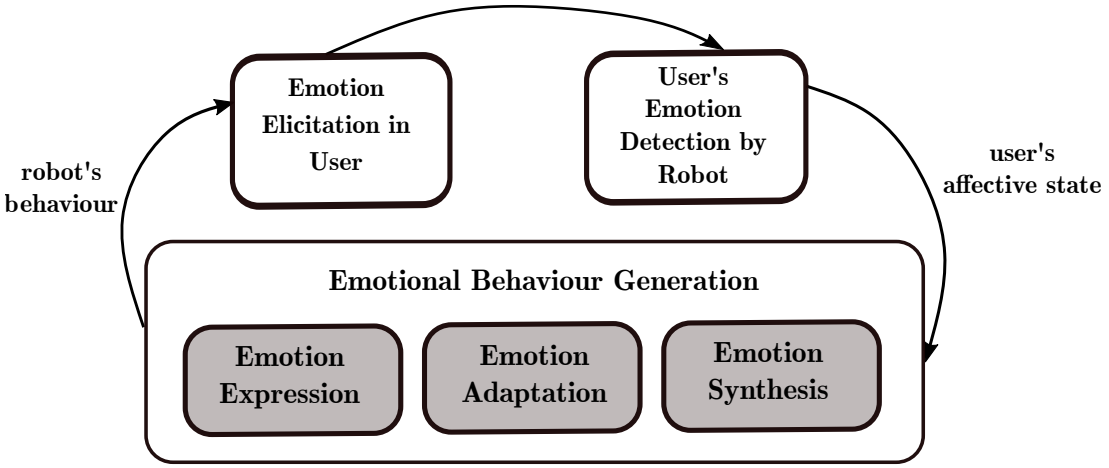

Fig. 1.1 The affective loop experience in human-robot interaction. Image recreated from the original in Paiva et al. [180].

1. Give the illusion of life, i.e., the robot's affective response should be believable to increase its social presence.
2. Augment engagement in a social interaction context.
3. Augment social presence in the long-term, by way of adaptive and personalised behaviours.

1.5 Scope, objectives and structure

This section will attempt to pinpoint our contribution concerning the general context presented above. Our work is primarily inspired by the affective loop schema for affective human-robot interaction. The assumptions we apply in our approach is that the user affect detection, and the three modules of emotion behaviour generation (emotion synthesis, emotion adaptation, and emotion expression) can be implemented independently and communicate through a common numerical representation of emotion as their input and output.

We decided to assume the independent development of each module because each functionality (i.e., detection, synthesis, adaptation and expression) is highly complicated on its own, intertwined with many different theories, possible models, and architectures, and thus, to develop the whole framework at once would require to consider too many parameters simultaneously. Such an approach would demand an enormous effort, and what is more, evaluating such a convoluted system could be very challenging because each module's

contribution might not be straightforward to assess. Based on these reasons, we decided to follow a modular approach.

Furthermore, we presume the availability of the emotion detection system, as well as the emotion synthesis and adaptation architectures, and we focus on the development of an emotion expression module. We assume the use of an emotion detection system that analyses several channels of human affect and outputs the classification result in the preferred emotion representation. This result is passed to the emotion synthesis and adaptation architecture, which subsequently outputs the synthesised emotion class of the robot, in the preferential emotion representation, adapted to the situation and the user's personality. Our goal is to build a framework that uses this output to generate the robot's emotional behaviour. Ideally, each module outputs a numerical representation of emotion and passes it to the next module, which modifies it according to its functionality. At the final step, this quantity is passed as input to the emotion expression module, which utilises it to generate the robot's behaviour. Regarding the emotion detection module, we assume the adoption of a commercial service, such as Affectiva[2]. For emotion synthesis and adaptation, a solution that could be adapted is the SIRE model [142]. For an overview of multimodal methods for emotion detection and emotion synthesis, the interested reader can access [150] and [180] respectively.

Furthermore, we constrain the scope of our proposed emotion expression module to nonverbal affect, or as we call it more often hereafter emotional body language (EBL), which constitutes our principal research interest. As we discussed in Section 1.3, there are several important modalities for EBL expression, but depending on the robotic platform, not all of them might be available. The robot we worked with for this project is Pepper, a humanoid robot created by SoftBank Robotics[3]. Pepper has a static facial expression; thus, we exclude this channel from our framework. The selected EBL modalities for Pepper in the scope of this project included body posture, motion, eye LEDs, and non-linguistic sounds. We excluded locomotion, ear and shoulder LEDs, and tablet display to reduce complexity, although these modalities are also available for Pepper. In particular, for non-linguistic sounds, although we included this channel in our preliminary analysis, we did not incorporate it in the EBL generation phase, because it would involve audio synthesis and generation techniques, which fall outside of the scope of this project since they are highly complicated tasks that would need to be tackled separately.

The work we present here focuses on the external aspect of emotion, i.e., emotion as a means of communication and social integration as discussed in Section 1.2, aiming into primarily enhancing the believability of a robot as a social agent in human-robot interaction.

[2]The Affectiva webpage: https://www.affectiva.com/

[3]The Pepper robot webpage: https://www.softbankrobotics.com/emea/en/pepper

The proposed framework can also be used for indicating internal states of the robot if these states are mapped to instances of the dimensional representation of affect. For example, battery levels could be scaled and passed to the EBL module as affect parameters, thus the outgoing expression would reflect the internal state of the battery, similarly to the work suggested by Lim et al. [142].

In Chapter 2, we discuss the predominant emotion representations and theories, along with their limitations. Subsequently, we present our choice, which is the dimensional model of core affect. The model is defined by two dimensions, valence (ranging from pleasure to displeasure) and arousal (ranging from activation to deactivation), and the values they take are numerical and continuous. We outline the reasons for our decision.

In Chapter 3, we review different approaches for robotic EBL synthesis. We focus primarily on humanoid robots, but we also discuss methods and evidence from studies on non-humanoids or embodied conversational agents. Literature review in this field is somewhat challenging to organise and present due to variations concerning the emotion representation models, the robot modalities and characteristics, the source of EBL used as a prototype, etc. Until recently, there were three main approaches for robotic EBL generation: 1) Direct human imitation with computer vision techniques using human EBL tracking, 2) Feature-based design, with human EBL observation, analysis and annotation to extract features and transfer them to the robot, and 3) Creative design, that is, conceiving and configuring key postures, and subsequently interpolate them as in graphical animation (pose-to-pose animation). We give examples from previous studies for all these approaches, and we discuss the challenges and the limitations in each case. Then, we present a more recent approach for robotic EBL synthesis driven by the progress and the enormous influence we witness in the field of artificial neural networks and deep learning. Our proposed framework follows this approach.

In Chapter 4, we present our initial experiment, which was designed to evaluate a set of robotic EBL animations created for Pepper by professional animators with the creative pose-to-pose method. We used this set of high-quality animations (context-free, nonverbal, dynamic expressions of emotion) because we assumed they could be of high believability and could serve as training examples for our generative algorithms. We conducted a user study ($N = 20$), in which participants evaluated the animations in terms of valence and arousal, the two dimensions of our selected emotion representation. We applied a reliability analysis on the collected ratings to aggregate the individual ratings into a single affect label of valence and arousal for each animation. Furthermore, we analysed the results to derive insights on human perception of robotic emotion expression.

Chapter 5 presents the deep generative model adopted for our proposed framework: the Variational Autoencoder (VAE). It is a probabilistic network that learns to reconstruct its input, and in doing so, it also learns a latent low-dimensional distribution that can be sampled after training to generate new data similar to the training examples. We begin with a brief discussion about deep generative modelling, and we continue with the standard autoenconder network, which, although not a generative model, bears similarities with the VAE architecture. The VAE framework is then described in detail as a directed probabilistic graphical model, and subsequently, from a deep learning point of view. Finally, we outline several challenges related to VAE training.

Chapter 6 presents our second experiment, the first application of the VAE framework for the generation of robotic EBL. As we noted before, early research in robotic EBL synthesis was mainly focused on hand-coded methods, e.g., pose-to-pose animation. This approach can result in high-quality animations, but since it is cumbersome and costly to design them, the animations are limited in number and granularity. This can be a drawback when we want long-term engagement on the user's side because the robot expressions will eventually appear repetitive and predictable. Drawing inspiration from deep learning methods, we propose to use a Variational Autoencoder network which is trained with a small set of EBL animations, created with the pose-to-pose method for Pepper (the ones we labelled in our first experiment presented in Chapter 4). In the sampling phase, we can generate numerous new animations, with many fine-grained variations. The generated animations are infused with salient features captured from the high-quality training examples, but they also contain emerging features. Furthermore, we explore the latent space of the model with a geometric approach, and we propose a feature that can potentially model the arousal dimension of emotion. The limitations and proposed next steps are outlined in the conclusions.

In Chapter 7, we tie everything together. We extend our EBL generation framework using a Conditional Variational Autoencoder (CVAE) for the generation of animations of targeted valence and arousal. This is accomplished by the generative process's explicit conditioning with the valence labels collected in the first experiment (Chapter 4), and implicit conditioning of the sampling with the radius feature proposed in our second experiment (Chapter 6).

In Chapter 8, we present the implementation of a new user study ($N = 20$) to evaluate the interpretability of the generated animations with respect to the valence and arousal conditioning.

Finally, in Chapter 9, we discuss the conclusions and limitations of this thesis's entire work. We also refer to some studies or applications that used parts of our work, and we briefly discuss some ethical considerations that could potentially arise from the deployment of robots using emotional body language.

In summary, based on studies that indicate the importance of affective body language in human communication [123, 229, 163], and evidence suggesting that its integration in robot behaviour can draw human attention [197], enhance user experience [172] and improve collaboration in HRI [137], our goal is to provide a pipeline for the automatic generation of robotic EBL expressions. Robots with the ability to mimic emotional expressions could be employed in various human-robot interaction settings and applications, be it human-robot collaboration tasks, robot-assisted therapy with kids or elderly, education, entertainment, public relation, etc. Previous approaches for EBL synthesis were primarily focused on hand-coded methods, which only produce a small number of animations, and thus limit the expressivity and believability of the robot in the long-term. Furthermore, many of the previous approaches use human EBL as a prototype and thus produce animations of a human-centric style of EBL, which may not be exaggerated enough to render the robot as a believable character (the illusion of life principle). Given that, we aimed to propose a data-driven, end-to-end alternative to hand-coded methods, a deep generative model that learns essential features from a small animation set originally designed with the robot morphology and the illusion of life principles in mind. Our work does not seek to be dismissive of hand-coded methods, on the contrary, we use their results as successful examples to train our network so that it learns from their features and create numerous new animations that will appear realistic and lifelike. Finally, although our work has been focused on humanoid robots, the methodology we propose could also be applied to other artificial agents such as non-humanoids or ECAs as long as there is an available set of animations in the form of joint angles per frame.

1.6 Contribution

The contributions of this thesis are the following:

1. We have compiled and labelled a set of 36 robotic EBL animations. These animations were originally hand-designed by expert animators for a Pepper robot. The labels, *valence* and *arousal*, are continuous interval values in $[0, 1]$. They are based on the dimensional representation of core affect, and they were aggregated from the ratings collected from the evaluations of 20 participants, after ensuring intra- and inter-rater reliability was high (Chapter 4).

2. In an exploratory analysis we examined how confidence in assessing valence in robotic EBL might be different than arousal (Chapter 4).

3. We designed and implemented an unsupervised pipeline using the VAE framework to generate new robotic EBL animations. The VAE was trained with the 36 robotic EBL animations, without the labels, and learned a latent representation which can be sampled, interpolated and decoded into new robotic EBL animations (Chapter 6).

4. In an exploratory study, we examined the latent space of the VAE, and we proposed a methodology to sample it systematically. We used this methodology to detect possible geometric features that could modulate the valence or arousal weight of the generated animations, and we derived the hypothesis that the radius of the latent space might be a candidate to modulate arousal (Chapter 6).

5. We extended the VAE-based pipeline so that it can generate robotic EBL of targeted valence and arousal content. The conditioning was accomplished by using the radius of the latent space to modulate arousal, while for valence we modified the network into a Conditional Variational Autoencoder (CVAE) which is trained and sampled with the valence labels (Chapter 7).

6. We conducted a user study with 20 participants to evaluate the interpretability of the valence and arousal conditioning of the CVAE-generated robotic EBL animations. Furthermore, we examined if participants perceived the generated and the hand-designed animations differently in terms of anthropomorphism and animacy (Chapter 8).

7. We released a set of robotic EBL animations that contains the 36 hand-designed animations we used for training the VAE and CVAE. We have recorded the data directly from a physical Pepper robot executing the animations. The set contains values for the 17 joints and 16 eye LEDs of a Pepper robot, as well as the valence and arousal labels[4].

The author of this thesis is exclusively responsible for writing this manuscript, the publications mentioned in the Author's Declaration and the code for implementing the algorithms, running the experiments, analysing the data (whenever other packages have been used, they are explicitly mentioned in the Software sections of each chapter). The second user study described in Chapter 8, was conducted by Fernando Garcia, but the experimental design, the implementation of the interface, and the data analysis have been exclusively produced by the author of this thesis. The use of plural first-person pronouns in this manuscript reflects the appreciation to several people who offered their ideas, advice and feedback (please refer to the Acknowledgements section).

[4]REBL-Pepper Dataset: https://github.com/minamar/rebl-pepper-data

Chapter 2

Emotion representation

2.1 Introduction

This chapter will present the emotion representation approach we adopted for our experiments and technical implementations. More specifically, this choice applies to the labelling, annotation and interpretation of robotic EBL animation concerning our user studies. It is also a fundamental component in implementing the deep learning framework for the automatic generation of robotic EBL, which we will present in Chapters 6 and 7.

Several emotion representations have emerged from prominent psychology theories and models of emotion, but after many years of research, there is no consensus on which approach is the most appropriate to explain and model emotions. This problem arises mainly from the fact that we cannot measure emotions directly and objectively. We can only use introspection or outward expression evaluation to analyse them, which are both highly subjective methods, and thus emotion is ill-defined [135].

The rest of this chapter is divided into two sections. In the first section, we discuss the categorical and the dimensional representation of emotion, their respective psychological emotion theories and limitations. At the end of this section, we also discuss the appraisal models of emotion briefly. In the second section, we explain our decision to adopt the dimensional approach.

2.2 Emotion representation paradigms

There are three representations of emotion classification and modelling prevailing in affective computing and social robotics. The *categorical representation*, which is based on the discrete basic emotions theories and the locationist approach, the *dimensional representation*, backed

up by the dimensional models of emotion and the psychological constructionist approach, and the *appraisal theory of emotion*, in which emotions arise from individual evaluations applied on events under the light of a person's personal beliefs, desires and intentions.

2.2.1 Categorical representation of emotion

According to the categorical representation, to classify different affect states, we can use a few categories corresponding to basic emotions. The categorical approach is based on the theory of discrete basic emotions which posits that there is a core set of distinct emotion categories, the basic emotions, each corresponding to distinct brain locations or patterns of brain activity (hence it is also called locationist approach), each manifesting with a specific feeling and physiological pattern [187]. These primary emotion modules are considered as biologically basic in the sense that they cannot be segregated further in simpler building blocks. Still, they can be combined to give rise to higher complexity emotions, and perhaps even exhibit mechanisms of emergent properties [181]. Proponents of the theory of the basic emotions uphold the universality of this multimodular emotion system which they postulate is shared within the human species, and even with some other primates [66, 181]. Furthermore, they maintain that basic emotions are principally prewired responses to different stimuli, an "affect program" written by the evolutionary process through natural selection, which can be somewhat influenced epigenetically by learning.

In his pioneering work, Paul Ekman, one of the most prominent advocates of the theory of basic emotions, sought to demonstrate the existence of a facial affect program that is innate and cross-cultural, and when activated by some elicited primary emotion, it triggers specific and universally common facial muscular movements [65]. This triggering of facial expressions is largely involuntary, although it can be controlled at some extend by learned habits (display rules), resulting in some visual changes on the distinct pattern of the facial expression. In an attempt to further systematising this theory, Ekman and Friesen published in 1978 their Facial Action Coding System (FACS), a method that can be used to describe and measure facial movement based on an anatomical analysis of facial action [68]. In a similar vein, coding schemes have been proposed to describe the bodily expression of emotion [56, 102, 103].

Besides being very influential, the theories of discrete basic emotions have also been criticised on various grounds. As we explained before, a locationist account of the brain basis of emotion postulates that every primary or basic emotion is localised to a specific brain locale or anatomic network. Lisa Feldman Barrett challenges this notion by scrutinising findings from two meta-analyses of emotion neuroimaging studies which fail to provide reliable and robust evidence of unambiguous localisation of basic emotions in the human brain, while in

some cases, the same brain areas are found to be activated by different basic emotion stimuli [10, 13]. Lindquist et al. [144] in their meta-analytic review of the human neuroimaging literature on emotion, conclude similarly that the evidence for the consistent and specific localisation of basic emotions in brain regions are not sufficient enough not to be considered incidental. They find more likely the presence of a highly complex interplay between brain regions and anatomical networks, detectable in both emotional and non-emotional experiences (e.g., language, executive control, action simulation), and thus non-specific to emotion.

Barrett in [10] also challenges the hypothesis according to which each basic emotion manifests with a specific physiological, behavioural or facial response, and these responses are correlated. She maintains that the evidence reveals a lack of response coherence within each emotion category, and dismisses the assumption that one modality can be used as a proxy for all the others.

Furthermore, Barrett rejects the notion of emotion-specific bodily response whether this is voice, facial expression, or body posture because those channels do not necessarily display information about the person's emotional state, e.g., people can feel happy without smiling. Most importantly, the emotion-specific bodily response assumption is also objected on the grounds of universality, a principal claim in Ekman's theory, lack of which would point to lack of emotion-specific and consistent physiological or facial response. Gendron et al. [81] support this objection with a study comparing emotion perception of facial expression between participants from the US and the Himba ethnic group, a sample from a remote cultural context with limited exposure to Western culture. The findings support the view that people from different cultures have different notions of emotional expressions, even when linguistically relative emotion terms exist in their languages.

Finally, Barrett's criticism also points towards methodological problems. For example, in Ekman's emotion recognition tasks, the perceivers of emotional facial expressions are forced to select a label from a small list of basic emotions, something that is thought to inflate agreement.

Currently, the categorical paradigm has been the most widely adopted in the field of affective computing [90]. Several different lists of basic emotions have been proposed, but arguably the most frequently used one in emotion recognition tasks is Ekman's list of six basic emotions: anger, disgust, fear, happiness, sadness and surprise. However, there is growing concern about the limitations of categorical models on account of the loss of information associated with the effort to classify rich emotional states of high complexity with a handful of emotion categories [90]. These labels can be potentially very restrictive for automatic emotion perception or modelling in artificial agents or even induce stereotypical

and biased judgements, perhaps with implications if we consider sensitive applications related to security, mental health etc.

2.2.2 Dimensional representation of emotion

The dimensional models define the affect states according to one or more dimensions, with pleasure and intensity being the most common choices. The psychological constructionist account of emotion influences dimensional models according to which, emotions are psychological events, and they emerge from basic psychological operations which are not specific to emotion [144]. Emotions do not have a particular fingerprint, e.g., smile when happy, but they are constructed on the spot by general-purpose brain networks and functions responsible for multiple cognitive processes.

According to the theory of constructed emotion [12]:

> In every waking moment, your brain uses past experience, organised as concepts, to guide your actions and give your sensations meaning. When the concepts involved are emotion concepts, your brain constructs instances of emotion.

Constructionists argue that emotions are *created* by the mind out of our sensory input, bodily sensations, situational information, and prior experience [144]. They are not just involuntary reactions to the world, because humans are not just passive receivers of sensory information from the world [12]. To the contrary, we actively construct our emotions in the same way we create meaning out of sensory input and prior experience. In particular, the sensory input and bodily sensations and their neurophysiological correlates, define the *core affect* [196, 195], i.e., the most elementary affective feelings of hedonic pleasure or displeasure with some degree of activation or deactivation.

Core affect can be perceived consciously, is always present, but it does not need to be directed at or attributed to something. It can be experienced as a free-floating mood, simply feeling good or bad, aroused or relaxed. But when it becomes directed (attributed to some cause) [195], or when it is categorized as a *situated conceptualization* given a physical context or a context-specific memory [11], then it becomes part of a prototypical emotional episode. The prototypical emotional episode is the full-blown emotion, which encapsulates core-affect but also involves more complicated cognitive functions related to directionality (attribution) or situated conceptualisation (making meaning based on context).

The *circumplex model of emotion* [196, 195] represents core affect on a two-dimensional space defined by the valence dimension ranging on a continuum from pleasure to displeasure on the horizontal axis, and the dimension of arousal ranging from activation to deactivation on the vertical axis. Russel and Barrett [196] suggest that these two affective dimensions

may be sufficient to capture the core affect, which becomes a full-blown emotion when it is assessed within a situational framework. This representation allows us to depict each emotion (its core affect component to be precise) as a data point on this 2D space, which is visually intuitive and readily informative, and makes clear how different emotions are not independent, but instead, they are variations of the same two variables. Furthermore, this representation does not suffer from possible linguistic biases because valence and arousal are related to more basic psychological processes and not beliefs about emotion words or categories, which increases the validity of self-reports [11].

From a neurophysiological perspective, valence and arousal have been found in fMRI studies to correlate with activation of different brain circuits during emotional responses, which supports the validity of a two-dimensional view of the organisation of emotions [2]. However, as we have discussed so far, by definition, core affect can only be considered as less or equal to the full-blown emotion. Hence, a well-known criticism of the circumplex model of emotion is that the two affective dimensions of valence and arousal are insufficient for the differentiation between emotions that share common core affect, e.g., fear and anger both have negative valence and high arousal.

Several attempts have opted to solve this drawback by suggesting additional dimensions to improve the disentanglement of emotions with similar core affect. One of the most widely used third dimensions is dominance, which is also called potency, agency or control, and is described as the feeling of being in control, being able to influence one's environment. Mehrabian and Russell [164] initially proposed dominance in the PAD model, which uses three dimensions to represent emotions: Pleasure, Arousal and Dominance. It has been endorsed by multiple researchers as a third additional dimension to solve the problem of disentangling emotions with similar core affect, and evidence from neuroimaging studies seem to give some support although not conclusive [109].

Other dimensions have also been proposed. Davidson and Irwin [60] suggested approach-avoidance (or approach-withdrawal) as fundamental motivational states that give rise to emotional reactions and an organism tendency to approach or avoid a stimulus. However, some researchers consider that valence might be correlated with this dimension since positive valence is associated with approach motivation and negative valence with withdrawal [93]. Fontaine et al. [74] claim that to sufficiently represent similarities and differences in the meaning of emotion words, four dimensions are required, and they suggest the following ones in order of importance: evaluation-pleasantness, potency-control, activation-arousal, and unpredictability. Appraisals of novelty and unpredictability characterise the last dimension.

These arguments highlight several researchers' agreement on the need for additional dimensions to describe full-blown emotions, but clearly, there is no consensus on which ones

to employ. In any case, the optimal number of extra dimensions depends on the research questions and goals.

2.2.3 Appraisal models of emotion

In the core of the appraisal theory of emotions lies the *person-environment relationship* [133]. Emotions are the results of subjective evaluations (appraisals) of external stimuli (events) filtered by the internal state (goals and beliefs) of the subject experiencing them [179, 200]. Events elicit emotion episodes bounded in time, i.e., they have a clear onset and a less distinct offset. Appraisals are applied continuously, in recurrent and hierarchical cycles, resulting in appropriate action tendencies that serve adaptive functions. They are cognitive but not necessarily conscious.

According to appraisal theorists, particular emotions are linked to specific appraisals, but the same event can trigger multiple appraisals. Thus, an event is considered as less definitive compared to the appraisals triggered by it. The full-blown emotion has several components: elicitation processes, physiological signature, motivation, motor responses, and subjective feeling [199].

Computational appraisal models focus on the appraisal as the primary process, and they encode elaborate mechanisms to derive the appraisal variables. Appraisal variables are associated with individual judgements used by the artificial agent to generate an emotional response. On the other hand, the emotion itself is modelled in a far simpler way, often just with a label or intensity, which is derived from the appraisal variables with *if-then-rules* [87].

Appraisal theories have been very influential and broadly used in affective computing, mainly in models investigating the impact of emotion in cognitive and behavioural processes. One of the criticisms on the appraisal theories questions the assumption that cognitive processes always precede the affective reactions. Arguably, these two processes could also occur in reverse order, and what is more, emotions could also be triggered by hormones, i.e., factors that have nothing to do with judgements and cognition [195].

2.3 Emotion representation in our work

The appraisal theories propose a cognitive framework to model emotion. In this model, signals received by the environment and contextual information are processed by an agent's system of beliefs, intentions and desires, to produce an emotion, the emergence of which triggers a behavioural or cognitive response, often related to decision-making. Our objective is to use emotion representation as a given descriptor of the outward, context-free robotic

expression of affect, merely as a label to be used for the automatic generation of robotic EBL. Therefore, the agent in our work is not equipped with a context processing component or with mental attitudes (i.e., beliefs, intentions or desires) which would enable the formation of appraisals to trigger an emotion label, because the latter is a given in the framework we propose. Our robotic EBL generation framework could perhaps be used as a component of an appraisal model. Still, using an overly cognitive architecture to represent emotion as a descriptor label in our framework would add considerable complexity which would outweigh the advantages.

As far as the two other representations, the categorical and the dimensional model, significant efforts have been made to compare their benefits and limitations, [101, 55, 36, 171]. Nonetheless, the question of how to select an appropriate model for a given problem remains open. In our present study, we decided to adopt the dimensional representation and the circumplex model of emotion. Although this approach is not general enough to model the full-blown emotion, it can provide a fair representation of non-directional core affect, a valid low-dimensional component of the full-blown emotion [145].

Low dimensionality was a key factor for our choice for two reasons. First, we want to gradually scale our representation's complexity, starting from the more straightforward set of components, i.e., valence and arousal, evaluate it and then add more complexity. Second, since we will use self-reports in user studies, we want to keep the participants' workload as low as possible to retain their motivation for several trials. Therefore, although we are focusing on valence and arousal for the technical part of the work presented in this thesis (Chapters 4, 6 and 7), in our final user study presented in Chapter 8, we also include the dimension of dominance, as a third evaluation component. This is an exploratory step towards collecting additional information on how humans perceive robotic EBL, hoping that this information can provide suggestions for future work on extending our emotional representation to involve another dimension.

Another decisive factor for our choice to work with a dimensional model is that it permits us to describe an emotion using continuous numerical values. Since our objective is to work with deep neural networks to synthesise robotic EBL, using labels or features that are numerical and continuous is mathematically convenient. With categorical strings as labels, we would have to use encodings (e.g., one-hot encoded or dummy encoded vectors) since most machine learning algorithms take numerical values as input. This is a credible solution, but of higher computational complexity because the vector representation size grows with the size of the word corpus, and it is also very sparse (a single 1-element and a lot of zeroes) which makes the computation more costly and slow. Additionally, thinking of future applications in which robot emotion expression is coupled with multimodal emotion recognition, working

in a continuous and highly discretised space would be more intuitive and convenient from an engineering point of view, since fusing or averaging numerical features from different modalities is more straightforward compared to fusing emotion words.

Besides the technical advantages of working directly with continuous numeric features in deep neural networks, we wanted to avoid working with categorical emotion words because of possible linguistic bias arising from individual differences in the semantic interpretation of the emotion words during the labelling phase. Furthermore, if we had decided to proceed with the categorical representation, another challenge would have been choosing the emotion words list. The most obvious choice would have been Ekman's six basic emotions, since adding more words would significantly increase the participants' workload. However, the shorter the list is, the more likely the participants are to select labels that do not necessarily reflect their beliefs due to the forced-choice response format [194].

To conclude, we are convinced that following a dimensional approach will benefit our project, both in the EBL labelling phase, and the technical implementation phase. In the first phase, we will be able to derive affect labels of valence and arousal in a continuous and highly discretised space; thus, these labels will be potentially more effective in capturing different subtle variations of the emotional state than categorical tags that aggregate many expressions in a small number of categories. In terms of the technical implementation phase, where we aim to build a deep learning framework for generating new emotional robotic expressions, this approach will reduce complexity and improve stability. Furthermore, training a network with higher sensitivity and variability labels has more potential to inject these properties in the output and generate expressions of fine-grained differentiation. Summarising, we will use a dimensional approach for our core affect representation, with the affective dimensions of valence and arousal.

Chapter 3

Related work in robotic EBL synthesis

3.1 Introduction

Affective body expressions are essential for conveying and perceiving emotion among humans [123] and human-robot interaction. In particular, for humanoid robots with constrained facial expression, other channels such as body posture, LEDs and sounds can be used to express affect. The problem of designing bodily expressions of emotions in humanoid robots has attracted much attention in recent years, and several efforts have been made to design effective expressions and test their impact on humans. However, it is difficult to juxtapose different efforts due to differences in the design principles adopted, the robot's level of embodiment, the modalities involved in the expressions, the static or dynamic nature of the expressions, the emotion representation model, and the evaluation methodology. Early research in this direction can be summarised in three main approaches: 1) Direct human imitation, 2) Feature-based design, and 3) Creative design. In this chapter, we will outline and discuss previous work following each approach. In some cases, researchers might choose to infuse aspects from one approach to another. A fourth approach using deep learning has appeared very recently, and we discuss it last since it is the one we will also follow in our work. Our literature review is focused predominantly on work which reports user studies evaluating emotion recognition or impact instead of other measures. Our review is focused mainly on humanoid robots. However, we also discuss selected work on non-humanoids and embodied conversational agents. For a broader systematic review on robotic animation techniques that includes additional robotic designs, evaluation metrics and expression channels, the interested reader is directed to Schulz et al. [201].

3.2 Direct human imitation

In the first approach, the robot joints are configured to match the human joints in a single posture or motion setting. To accomplish this, computer vision techniques, markers, or sensors are employed to track key joints positions in human body motion, and then map them to the robot joint space manually by observation [124, 21], or by way of a transfer function [158]. Regarding the second method, although it has an advantage as an end-to-end process, constructing useful transfer functions can be very challenging. There are no studies so far applying it specifically for robotic EBL synthesis to the best of our knowledge.

Beck et al. [19, 21] designed six poses to convey the six basic emotions on a Nao robot. For each pose, the robot joints were configured to match by observation a human posture extracted as a key pose from motion-capture data. These data were recorded from a professional actor's performance of the corresponding emotion. The number of poses was further augmented by changing Nao's head position for each pose (looking up, down, straight). In their user study with a physical Nao robot, participants correctly identified Nao's key poses at better than chance levels. In the same experiment, valence, arousal and stance were also rated by the participants on a 10-point Likert scale, and it was found that the head position had an effect on these ratings for all basic emotions except *fear*.

3.3 Feature-based design

A prominent approach in robotic EBL design is to use features extracted from human EBL. Such features can be found in studies coding patterns of movement or posture from recordings of actors performing emotional expressions [58, 165, 56, 222, 224], computer-generated mannequin figures [53], or human motion capture data [122]. These studies provide lists of validated features, such as body orientation, the symmetry of limbs, joint positions, force, and velocity, which can be employed to design robotic EBL expressions.

Häring et al. [105] designed a set of multimodal (body motion, sound and eye LEDs colour) animations for a Nao robot, to express anger, fear, sadness and joy. In the case of body motion, they wanted to approximate human behaviour using Coulson's [53] static postures and de Meijer's [165] gross body movements features. Regarding the sounds, the choice was of human or animal-like acoustic expressions that are "commonly related" to human emotions as the authors reckon (e.g., crying for sadness). Eye LEDs colours were selected similarly (e.g., red for anger). All three modalities were tested separately. The evaluation was dimensional, including pleasure, arousal and dominance (the PAD model [164]). The results did not show any significant effect resulting from the eye LED colour

modality, and only half of the sounds used were categorised appropriately. However, all the body movements were rated in the right octant of the PAD model except *sadness* which was rated with positive arousal.

Embgen et al. [69] designed EBL animations for a Daryl robot to express three basic emotions (happiness, sadness, fear) and three secondary ones (curiosity, embarrassment, disappointment). Their design principles for the body configurations were based on proposed features from Darwin [58], Wallbott [222], and Gunes and Piccardi [89]. For the colour of the chest LED plate, they were inspired by idiomatic references to colours (e.g., "feeling blue"). Regarding their user study, they reported that participants could identify the emotions from the EBL displayed on a physical Daryl and distinguish between the somewhat similar emotions sadness and disappointment.

Erden [70] designed EBL postures for a Nao robot based on Coulson's [53] static postures. Their designed postures aimed to express three emotions: anger, sadness, and happiness. In their user study, participants were presented with photos of the Nao postures. The report that *anger* was recognized with 45%, *happiness* with 72.5%, and *sadness* with 62.5% after .

Destephe et al. [62] extracted features from motion capture data recorded from two actors who improvised on emotional straight walking. The emotion categories used were sadness, happiness, anger, fear and neutral. For each category, there were also four categories of intensity: low, intermediate, high, and exaggerated. The extracted features included step height, step length, velocity, cadence, head pitch, shoulder pitch and waist pitch. The features were subsequently adjusted to fit the WABIAN-2R humanoid robot's kinematic structure, and the walking patterns were displayed on the virtual robot. The evaluation was conducted with the recorded videos of the virtual robot. The authors report a high recognition rate for the emotions (72.32% average for all emotions, *fear* was the highest). The intensity recognition rate was not as high (33.63% average for all intensities, *intermediate* intensity was the highest), but they found that it modulates the emotion recognition rate (higher intensity results to higher emotion recognition rate).

Tsiourti et al. [214] designed emotional animations for a Pepper robot and a Hobbit robot to test the interpretability of three basic emotions (happy, sad and surprised). The robotic animations for Pepper were created by selecting expressive postures from Thomas and Johnston [79], configuring the joints of the robot to match postures proposed by Coulson et al. [53] for each emotion and subsequently using features such as velocity and amplitude proposed in Kleinsmith et al. [122] to derive robotic animations from the postures. The evaluation was conducted on videos of the recorded robot. *Happiness* was recognized more accurately, while *sadness* and *surprise* were poorly recognized from body motion alone.

McColl and Nezat [159] used features from [222, 165] to design EBL for the humanoid Brian 2.0. The emotions that sought to convey were sadness, elated joy, anger, interest, fear, surprise, boredom and happiness. The expressions involved the upper body of the robot. The evaluation was conducted on videos. Sadness and surprise had the highest recognition rate ($> 80\%$), while fear and happiness had the lowest recognition rate ($< 30\%$).

An interesting paradigm following the feature-based approach is inspired by the Laban Motion framework [131], based on kinesiology, anatomy, and psychological analysis of human motion. Often, researchers use only a subset of the framework, the Laban Effort System, which provides four motion parameters: space, weight, time, and flow. These parameters have been used to handcraft features and generate locomotion trajectories that express affect in non-humanoid robots with low degrees of freedom [125, 203, 3, 176, 211], with promising results in terms of readability by humans. For example, Sharma et al. [203] employed motion capture to record human emotional motion, which they subsequently decoded into space, weight, time, and flow values with the help of a Laban-trained artist. Then, they used the values to configure a set of flying patterns for a Parrot AR.Drone. For the evaluation, the physical robot executed the flight patterns. The participants were asked to assign rates of valence and arousal. The researchers found that space and time were significant predictors of valence, while all four Laban parameters could predict arousal with statistical significance. Takahashi et al. [211] designed head and arms motions to convey the six basic emotions for a teddy bear robot using the Laban principles. Their evaluation showed that although fear and disgust had a low recognition rate, the rest of the emotions were well recognised. The Laban Motion framework has also been used to design EBL for humanoids [157, 156, 155, 175, 174]. Masuda et al. [157] used Laban features (space, time, weight, inclination, height, area) to modify three basic movements for the humanoid robot HFR to express emotion. Their evaluation shows successful rates of recognition for some of the modulated movements. Concerning the emotions, *sadness* appeared to have the highest modulating effect.

Another interesting approach is to decode human voice features into robot motion. The SIRE model [142] was proposed for generating robot expressions of happiness, sadness, anger and fear, based on four features: speed, intensity, regularity and extent. These features were extracted from human expressions, in particular vocal expressions. The model was evaluated on a Nao robot using the dimensional PAD model, and the results showed successful recognition of *happiness* and *sadness*. The SIRE model was also adapted in another study to produce expressions of emotions on a Nao robot that would offer movie recommendations [191]. This Nao was found preferable and more persuasive than a Nao without emotional expression.

Feature-based design can also be inspired by animal models instead of human EBL, an approach which can be potentially more effective with pet-like or animal-like robots [166]. Lakatos et al. [132] looked into the effects of robotic EBL inspired by functionally analogous dog behaviours. They used a MogiRobi robot, a dog-like mechanical robot which moves on a wheeled base and can move its head, ears, and a rear-antenna. They modelled the emotions of fear and joy. The assessment showed that participants recognised the happiness expression very well, but fear was less clearly perceived, and often mistaken as indifference or not an emotion. Animal-like features for robot expression have also been tested with Probo [198], an elephant-like robot, and even a non-animal robot head, EDDIE, [130]. However, both studies have mainly focused on emotion modelling based on facial expression. Other interesting studies on emotion modelling for animal-like robots include the seal-robot Paro [204, 220], the dog-robot AIBO [212], the cat-robot NeCoRo [140]. Although these studies have not directly evaluated emotion recognition and interpretability, they provide some indirect evaluation by testing the effects of using these emotionally expressive robots on robot-assisted therapy.

3.4 Creative design

Another influential approach emerges from applying the artistic principles and practices used in graphc animation to generate robot expressions of emotion. The animator conceives several expressive key postures with a creative approach and the robot morphology in mind, configures the robot according to these postures, and then applies the *inbetweening* process, i.e., creates the intermediary postures so that the sequence appears continuous and smooth. This is the pose-to-pose animation technique. This procedure is robot-centric and often influenced by Disney's twelve basic principles of animation [79], producing very lifelike expressions. In some studies, puppeteers [139, 213] or random participants [230, 108] are enlisted to configure robot EBL postures or sequences of postures according to their subjective perception.

Monceaux et al. [169] describe their methodology of creating a big library of dynamic multimodal emotion expressions as an iterative process involving imitation from direct observation, and abstraction of key positions to adjust them to the Nao robot morphology. The key postures were inbetweened with interpolation to produce a robotic animation (pose-to-pose animation).

Ribeiro and Paiva [188] adapted Disney's twelve basic principles of animation [79] for an EMYS head robot to express the six basic emotions, on three different intensities. Their

user study evaluation found that *anger*, *sadness* and *fear* were very well recognized, surprise and joy were fairly recognized, but *disgust* was poorly identified for low intensity.

The principles of animation have also been used to design affect expression for non-humanoid robots. A fascinating example is the affect display designed by Yohanan and MacLean [231] for the Haptic Creature, an animal-like robot. Their methodology followed a pose-to-pose design of motions for the robots ears, lungs and purr. The goal was to create nine different expressions as a combination of three valence levels (negative, neutral, and positive), and three levels of arousal (high, medium, and low). The evaluation showed the design was effective in conveying arousal but ambiguous in the communication of valence.

3.5 Deep learning approaches

Lately, there have been some efforts to create body language using deep learning frameworks. Yoon et al. [232] implemented a co-speech gesture generation framework that uses an encoder-decoder network to produce mappings from sequences of words to sequences of a NAO robot poses. Their network is trained with human gesture and speech data. In the affective robotics field, Suguitan et al. [210] proposed the use of CycleGANs to generate affective robotic movements after training them with human EBL. They trained three different networks to generate animations expressing happiness, sadness, and anger for a Blossom robot with four degrees-of-freedom. They evaluated their model with a clustering inspection of the principal components of 13 different features extracted from their generated robot movements, and they report the intended emotions are fairly discernible between the three classes.

3.6 EBL synthesis in Embodied Conversational Agents

Generation of nonverbal emotional responses has attracted much attention in the area of Embodied Conversational Agents (ECAs) too. Nonetheless, since ECAs can have animated faces that allow elaborate expressions, facial display of emotion has been more widely exploited. In contrast, EBL has been approached in a more restricted way, predominantly as a co-verbal gesturing function that aims to accompany and be synchronised with speech [47].

In designing expressive motions for ECAs, a common approach is to begin with human motion data captured in conversational settings and then proceed with either a rule-based or data-driven generative strategy [78]. Rule-based generation exploits knowledge from psychology literature and perceptual studies to derive mappings from emotional states to facial expressions and body configurations, an approach similar to the feature-based design in robotic EBL described in Section 3.3. Data-driven generation uses corpora containing

segments of recorded human motion coupled with semantic information, similar to the direct imitation approach used in robotics EBL described in Section 3.2. The second strategy has an advantage: it allows for more sophisticated personality and stylistic effects to be injected in the overall motion, since the corpora can include a great variety of motions, as opposed to rule-based approaches that tend to average out individual differences [78]. Following, we will briefly discuss some research projects following one of these approaches, or often, a mixture of both. However, our discussion will not cover the evaluation of these ECA projects, since their evaluation is not targeted on assessing aspects of EBL, but follows a more holistic perspective with measures such as conversational smarts, smoothness of interaction and language skills. The interested reader is referred to the cited bibliography and a survey discussing several ECAs evaluation studies [76].

Justine Cassell with her pioneering work on ECAs [47] introduced the rule-based approach for the autonomous generation of verbal and nonverbal conversational behaviours in Rea architecture [46, 44, 45]. Rea is an ECA that acts as a real estate salesperson interacting with users and showing them around virtual houses. Rea's body language involves head nodding, hand gestures, facial expressions and other nonverbal communicative behaviours, mapped to communicative functions such as conversation initiation or termination, turn-taking, and salutation.

In the GRETA ECA system [183], the model of nonverbal behaviours has been influenced by psychology studies on human body language. From a large set of parameters proposed in perceptual studies [222, 223], a subset of six were selected and implemented in GRETA [94, 183, 27]. The six expressive parameters are: spatial extent, temporal extent, power, fluidity, repetition, and overall activation. Spatial extent corresponds to the amplitude of a motion signal, while the temporal extent is related to movement speed. Power corresponds to motion acceleration, and fluidity defines how smooth the transition between two successive movements is. Repetition refers to the number of times a movement is replayed, and overall activation defines quantitatively the overall amount of movements injected in a behaviour. The six expressive parameters have also been used to develop distinctive agents, each endowed with an individual personality baseline [149].

Following the data-driven approach, MAX [22], an agent powered by the WASABI appraisal model aims to directly map emotional states to specific gestures, instead of using expressive parameters to modulate motion. WASABI generates an emotional state which is subsequently mapped to a discrete emotion using the PAD space [164], and this mapping is used to select an appropriate movement from a predefined corpus of high-level abstract descriptions of the gestures.

Another data-driven approach example is the Physical Focus model [154, 86] which uses annotated motion capture segments to design EBL for virtual humans. The architecture uses the virtual agent's emotional state as input to a finite-state machine that determines a focus mode corresponding to a subset of nonverbal behaviours. The architecture uses four focus modes (strong body focus, body focus, transitional, and communicative) to group behaviours. For example, body focus mode represents self-focused attention that takes a distance from the conversation and is triggered by emotional states such as guilt or depression. In body focus mode, the behaviours involve gaze aversion, self-soothing or self-punitive gestures and minimal communicative gestures.

EMOTE (Expressive MOTion Engine) [7, 49, 233], is a 3D character animation system that uses the effort and shape components of the Laban Effort System to synthesise arm and torso movements. The idea of EMOTE is to apply the effort and shape parameters to underlying movements which were initially defined independently as pairs of time-frame and pose. The underlying movements can be of various origins, e.g., procedural generation or motion-capture transfer. By applying the effort and shape parameters, the final motions can become more natural and expressive.

3.7 Discussion on previous approaches

Previous research has been mainly focused on hand-coded methods for robotic EBL design. Hand-coded methods can be tedious and expensive, resulting in limited expressions, characterised by less granularity in their expressivity. Consequently, the expressions might appear repetitive and predictable, which might be disengaging in long-term human-robot interaction. Deep learning approaches seek to address this problem.

Another concern regarding the previous approaches is the common practice to use a human-like style of motion for the expression of a robot. Even if the emotion is recognisable, the robot expression might not be engaging enough in long-term interaction. As we discussed in Chapter 1, the illusion of life effect is essential in human-robot interaction if we want people to perceive robots as believable agents and engage with them [73, 137]. In our opinion, the believability effect is weakened when a highly complex signal, the human motion, is reduced to fit a significantly simpler robotic morphology. Scaling down the human motion range or excluding joints to match the robot's motion range may be practically simple, but it might miss important information or keep redundant and trivial information with a negative impact on the lifelikeness of the robot character. Similarly, automatic direct imitation methods suffer from a similar challenge [170]. Finding a good transfer function to transform the

robot's kinematics can be very challenging, even for androids with close similarity to the human body.

On the other hand, creative design methods involve human EBL only as an abstract source of inspiration for the animator, without posing strict constraints on the creative process [169]. In fact, the creative process could also be informed by other abstract sources, e.g., animal expression, graphical animation characters, arts or even subjective concepts. Combined with methodological principals such as Disney's twelve basic principles of animation [79] adapted to a specific robot morphology, as Ribeiro and Paiva [188] proposed, the robotic animations produced can be of high quality and very expressive.

3.8 Addressing the limitations

Our main objective is to propose an end-to-end method for the automatic generation of numerous new animations, which will appear smooth, natural and realistic, and exhibit increased granularity. To achieve that, we follow the deep learning approach, and more specifically, we use a generative model, the Variational Autoencoder.

As a starting point, we use a set of animations selected from Softbank Robotics animation library for the Pepper robot. Professional animators have created these animations to convey emotions, using the creative approach; after several expressive postures are conceived, the robot is configured accordingly, as a puppet or via the simulator, and the joints' angles are recorded. Subsequently, the postures are interpolated to get the intermediate states. With this choice, we wanted to take advantage of the professional animators' expertise to create a baseline set of advanced expressivity, smoothness and natural motion. This set can be used to train Variational Autoencoders, and we expect that the generated animations will be infused with the qualities of the hand-coded animations.

Lastly, another way in which we aim to differentiate from the majority of previous studies is in terms of the emotion representation, both in the labelling phase of the robotic EBL animations and the evaluation. We have already discussed in detail (Section 2.3) our choice to adopt the dimensional model of affect, with which we can avoid possible confounds related to forced-choice categorical emotion evaluation. Furthermore, our dimensional labels will be used as a discriminative marker and an integral component of the generative process.

In summary, the present thesis addresses the previous limitations as follows:

- We propose a deep learning methodology to achieve the **automatic generation** of an unlimited number of new robotic EBL animations, which are smooth, realistic and of fine granularity.

- Instead of using features extracted from human motion, we adopt a **robot-centric motion philosophy** in which we train the deep learning framework with a small animation set designed with the robot morphology and lifelikeness in mind.
- We fully adopt the **dimensional representation of emotion** in the labelling, learning and evaluation phases of our pipeline to benefit from the increased granularity it offers.

Chapter 4

Valence and arousal labels for robotic EBL animations

4.1 Introduction

The studies we presented in Chapter 3 are very informative for advancing robotic EBL synthesis because they propose and evaluate features and design principles regarding their impact on emotion readability. However, most of the studies focus on the categorical approach using the model of basic emotions, both in the design and evaluation phases. This is, of course, a legitimate choice since the debate between different representations of emotion is still open. Nevertheless, we believe that further study of the dimensional approach can also contribute to knowledge.

On the other hand, even in the case of studies using the dimensional models of affect at the evaluation phase, the evaluation is applied with coarse classes defined by a combination of two or three levels from each dimension (e.g., +/- Pleasure, +/- Arousal, +/- Dominance in [105]). The work presented in this chapter aims to take advantage of the properties of the dimensional model of affect, which due to its continuous-space representation facilitates granularity and fine-grained differentiation in the evaluation phase, even between expressions that only slightly vary. Essentially, our first study aims to define a reliably labelled set of robotic EBL animations, which can be used as the ground truth in user studies examining the impact of robotic emotional behaviours, or for designating outgoing robot behaviours in social HRI applications. Moreover, our labelled animations can be used as training examples for deep learning models, which is our ultimate goal in the overall project.

More specifically, we conducted an initial experiment to collect labels and knowledge that can help us to structure an emotion expression module. We used a set of high-quality

animations (context-free, nonverbal, dynamic expressions of emotion), designed by expert animators for the Pepper humanoid robot to convey emotions. We had them assessed in terms of valence and arousal in a user study ($N = 20$). The results provided us with insights into the human perception of robotic emotion expression. After performing a reliability analysis, we derived the affect labels of valence and arousal, in a continuous two-dimensional space for each animation.

To the best of our knowledge, this is the first effort to assign such labels to emotional body language animations for a humanoid robot, at least in terms of the reliability analysis and the increased discretization of the affect space. Robotic application developers could potentially use such a dataset to select outgoing emotional expressions based on continuous input signals, as a training set for deep learning models, or even as a ground truth from researchers that would like to study different setups related to human perception of robotic emotion expression. The dataset is released for other researchers to use it[1].

The rest of this chapter is organized as follows: the first section describes the user study set up, the participants, the robot platform, the animations, the interface used to collect data, and the experimental procedure. The second section outlines the reliability analysis and the results. The work we present in this chapter is published in [152].

4.2 Methods and materials

4.2.1 Participants

The study included 20 volunteers, 9 women and 11 men with a mean age of 23.6 years and a standard deviation of 4.09. The youngest participant was 19 and the oldest 35. The participants were mainly students at the University of Plymouth, UK. Nine participants were British, and the rest were European nationals studying in the UK. The participants had minimal or no experience with the Pepper robot. The experimental protocol was granted ethical approval by the University of Plymouth, Faculty of Science and Engineering Research Ethics Committee, and written informed consent was obtained from all participants before the study. The participants were reimbursed £5 for their time.

[1]REBL-Pepper Dataset: https://github.com/minamar/rebl-pepper-data

4.2.2 Robot platform and hardware

For this study, we used a physical Pepper robot, a humanoid created by SoftBank Robotics. Pepper is 120 cm tall, weighs 28 kg, and it has 20 degrees of freedom, including a head, two arms and a wheeled base. The operating system running on the robot was NAOqi 2.5.

4.2.3 EBL animation set

The set of emotional expressions used for this study consisted of 36 animations designed by SoftBank Robotics animators. Each animation has a different duration (from 1.96 to 10.36 s, mean = 4.39 s), and consists of body motion without locomotion. Some animations involved additional interactive modalities, i.e., eye LEDs colour patterns, non-linguistic sounds, or both. Essentially, each animation is a set of key body postures (frames) with their timestamps, and then intermediary frames are generated with interpolation. The interpolated frames are executed with a speed of 25 frames per second on the physical robot, but depending on the number of keyframes, each animation appears to have different speed. For a more detailed description of the animations, please refer to Appendix A, Table A.1, which presents the categorical tag of each animation assigned by the animators during the designing phase, a short description of the movements that comprise each animation, the duration in seconds, the speed calculated as total frames divided by keyframes, and the different modalities (motion, sound, LEDs).

Table 4.1 Nine classes of affect formed by 3x3 combinations of valence-arousal levels

		Valence		
		Negative	**Neutral**	**Positive**
Arousal	**Excited**	Negative/Excited	Neutral/Excited	Positive/Excited
	Calm	Negative/Calm	Neutral/Calm	Positive/Calm
	Tired	Negative/Tired	Neutral/Tired	Positive/Tired

The selection was made from the broader animation libraries for Pepper. We aimed to obtain a good spread of the animations across the affect space defined by valence and arousal. We predefined nine different classes of valence/arousal combinations presented in Table 4.1, and we selected four animations for each class. The selection was conducted based on the observation of the animations by the authors and a professional animator involved in the creative process of designing emotion animations for Pepper, i.e., researchers highly

accustomed to robot's motion and an expert in the expression of emotions with animations for the particular platform. Although we aimed to have a good representation for each class, the selection has been subjective, and the pre-assignment in the nine classes is not considered ground truth. Maximizing the affect space coverage with our pre-assignment has been a desirable property, and we look into how the preassigned classes match the final ratings. However, the final ratings provided by the participants were validated independently of the categorical tags.

4.2.4 Ratings interface

After carefully considering the available options for collecting valence and arousal ratings, we decided to use the *Affective Slider* [25], a modern digital tool that has been validated as a reliable mechanism for the quick measurement of valence and arousal. The Affective Slider enables the collection of ratings using continuous scales with customizable discretization. This study used scales from 0 to 1, with a step of 0.01, resulting in a resolution of 100 points. Furthermore, the Affective Slider is an intuitive interface that does not require significant training.

For our experiment, we replicated the tool based on the design guidelines recommended by its authors. We integrated it into the experimental interface along with a play button for the participants to control the robot, and a submit button to indicate that they finished configuring the sliders. The interface is presented in Fig. 4.1.

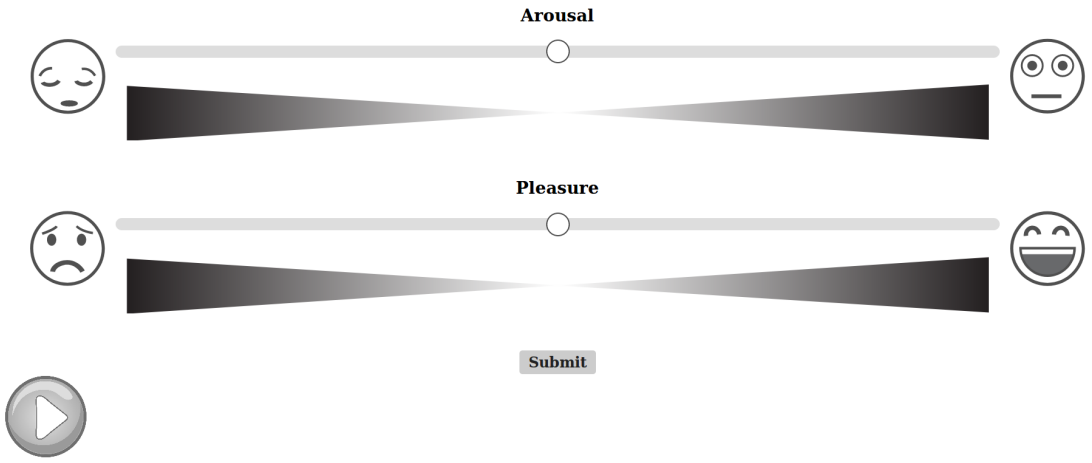

Fig. 4.1 The adapted Affective Slider interface for collecting valence and arousal ratings. Note that the valence slider is labeled as "Pleasure", to convey the concept of valence in a more intuitive manner for the participants.

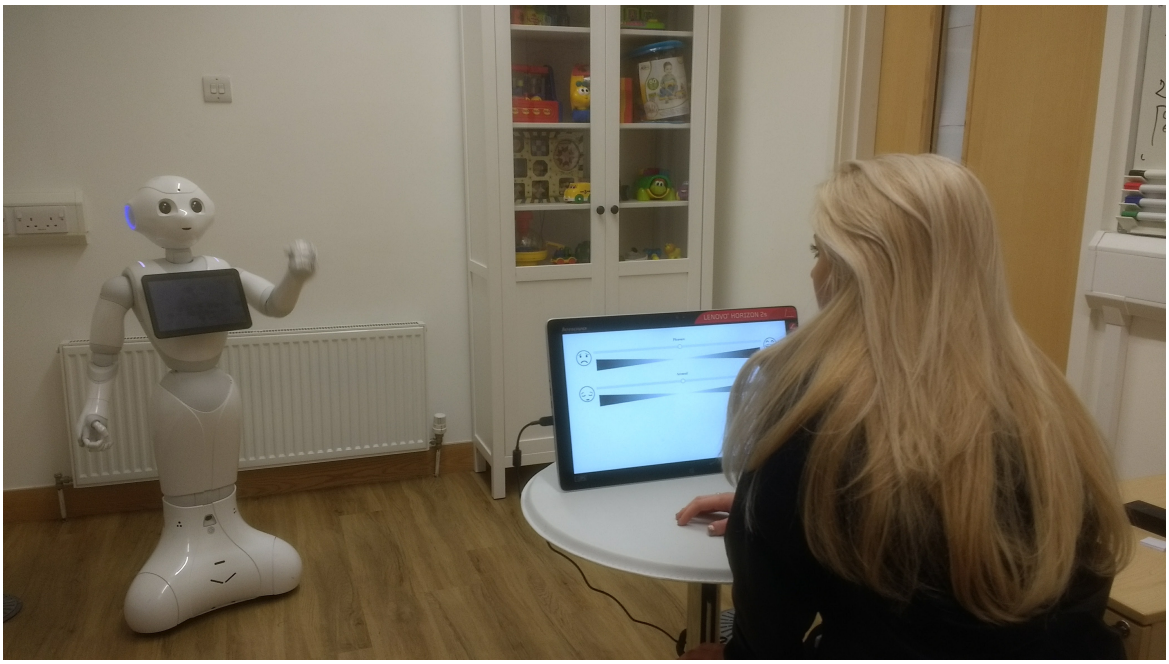

Fig. 4.2 The experimental setup.

4.2.5 Questionnaires

We administered the PANAS affect measuring questionnaire [225] ahead of the subject's interaction with the robot to indicate the subject's positive and negative mood at the time of the experiment. Furthermore, two 5-point Likert scales (from 1 = "Not confident at all" to 5 = "Very confident") were presented for the participants to indicate their level of confidence in their arousal and valence assessment after watching and evaluating each animation.

4.2.6 Experimental procedure

Instructions and training

Initially, the participants were provided with an information sheet and asked to give their written consent if they have no objections. After registering their demographic data (age, sex and nationality), they were seated in front of the robot at a distance of two meters with the touchscreen right in front of them. This setup is shown in Fig. 4.2.

The participants' first task was to complete the PANAS affect measuring questionnaire presented on the touch screen. Consequently, the experimenter gave a brief explanation of the main task. Participants were told that they had to rate the robot's emotional expression in terms of valence, referred to as "pleasure" during the experimental procedure, and arousal.

Valence was defined as "how positive, neutral or negative the robot's expression was" and arousal as "how excited, calm or tired the robot appeared". The participants' responses were solicited by the question *"How does the robot feel?"*. After this introduction, the participants went through a training session of three trials during which the experimenter was interacting with them to make sure they fully understood the task and the concepts of valence and arousal.

Main session

During the experiment's main session, the participants were presented with our adapted Affective Slider interface at each trial. First, they would click on the play button for the robot to perform the animation, and then submit their ratings. The participants were told that they could replay the animation if they needed to. When the participants clicked the submit button, the sliders would be removed, and two Likert scale questions would appear requesting them to indicate how confident they had been in their valence and arousal assessments. When these questions had been answered, the next trial would begin. The main session consisted of 39 trials; 36 original animations plus three repetitions. The order of the original animations was randomized for each participant to avoid order effects, and the first three animations were repeated at the end of each participant's session to evaluate intra-rater consistency. Moreover, the two Affective Sliders were presented in different order between trials. Between trials, the robot resumed its initial, neutral, motionless standing posture, with eye LEDs switched off, so that participants could perceive the onset and offset of the animation.

4.2.7 Software

For the execution of the robotic animations with the real Pepper robot we used NAOqi 2.5 SDK[2]. The participant's interface for controlling the robot and collecting the data was written in Python 2 (NAOqi is not migrated to Python 3) with Django v1.11.10[3]. For data preprocessing, wrangling and visualization, we used Python 3 and packages such as NumPy [178], SciPy [219], pandas [162], Matplotlib [104]. The statistical analysis of the collected data was carried out in R [186, 117, 80, 217].

[2]NAOqi 2.5: http://doc.aldebaran.com/2-5/home_pepper.html
[3]Django web framework: https://www.djangoproject.com/

4.3 Results

4.3.1 Affect questionnaire

The momentary affect scores from the PANAS questionnaire administered to the participants before the experiment were within the expected range [225], so following analysis did not exclude any of the participants. Positive affect mean was 31.75 ($SD = 6.27$) and Negative affect mean was 12.75 ($SD = 3.16$).

4.3.2 Descriptive statistics

The descriptive statistics by gender are presented in Table 4.2. An independent sample t-test did not reveal any statistically significant gender differences (valence: $t(18) = -0.62$, $p = .53$, arousal: $t(18) = -0.68$, $p = .49$).

Table 4.2 Descriptive statistics by gender. Ratings range from 0 to 1 with 100 points resolution.

Dimension	Gender	Mean	SD	Min	Max
Arousal	Women	0.60	0.29	0	1
	Men	0.61	0.27	0	1
	Group	0.61	0.28	0	1
Valence	Women	0.49	0.28	0	1
	Men	0.49	0.24	0	1
	Group	0.49	0.26	0	1

4.3.3 Exploratory analysis

In terms of the initial pre-assignment of the animations in nine different classes of valence/arousal, we plotted the means and the standard deviations of the collected ratings per class in Fig. 4.3, and the detailed descriptive statistics are presented in Table 4.3.

Since the animations in each class were subjectively selected as described in Section 4.2.3, there is no ground truth for the class centres (mean ratings per class). Interestingly, the figure shows that the class centres appear in the order expected according to the pre-assignment (see Table 4.1 for the order). However, almost half of the class centres are not spread out in the affect space as expected with the pre-assignment. For example, all the *Tired* class centres

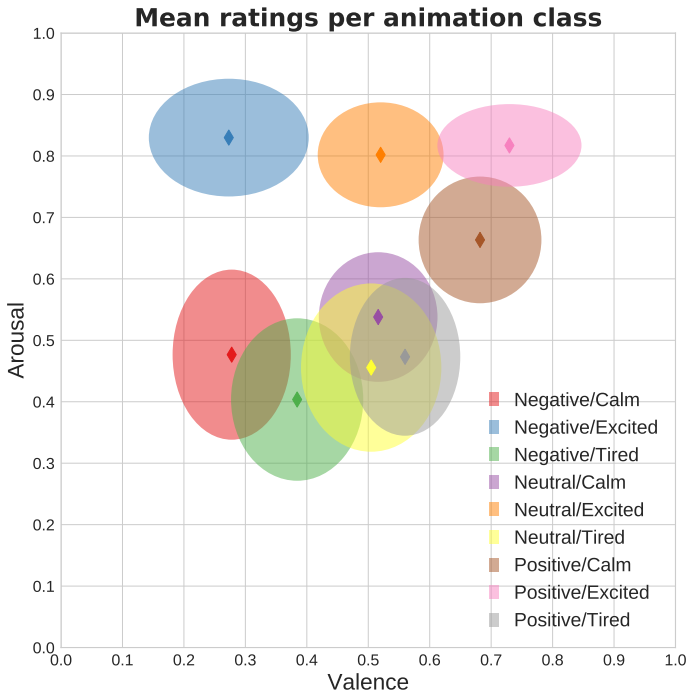

Fig. 4.3 Mean ratings and standard deviations of valence and arousal per pre-assigned affect class. The 2D space is divided into 9 classes as described in Table 4.1.

Table 4.3 Descriptive statistics by affect class of valence/arousal levels combinations.

	Valence		Arousal	
Animation Class (V/A)	Mean	SD	Mean	SD
Negative/Excited	0.27	0.26	0.83	0.19
Negative/Calm	0.28	0.19	0.48	0.28
Negative/Tired	0.38	0.21	0.40	0.26
Neutral/Excited	0.52	0.20	0.80	0.17
Neutral/Calm	0.52	0.19	0.54	0.21
Neutral/Tired	0.50	0.23	0.45	0.27
Positive/Excited	0.73	0.23	0.82	0.13
Positive/Calm	0.68	0.20	0.66	0.21
Positive/Tired	0.56	0.18	0.47	0.26

are pushed to the *Calm* subspace for arousal, and to the *Neutral* subspace for valence. The rest of the class centres appear in the expected levels of the affect space. It appears that the animations preassigned as of low arousal were in average perceived as of medium arousal

and their valence as neutral. For a visualization of the descriptive statistics of each animation in each class, see the boxplots in Fig. 4.4.

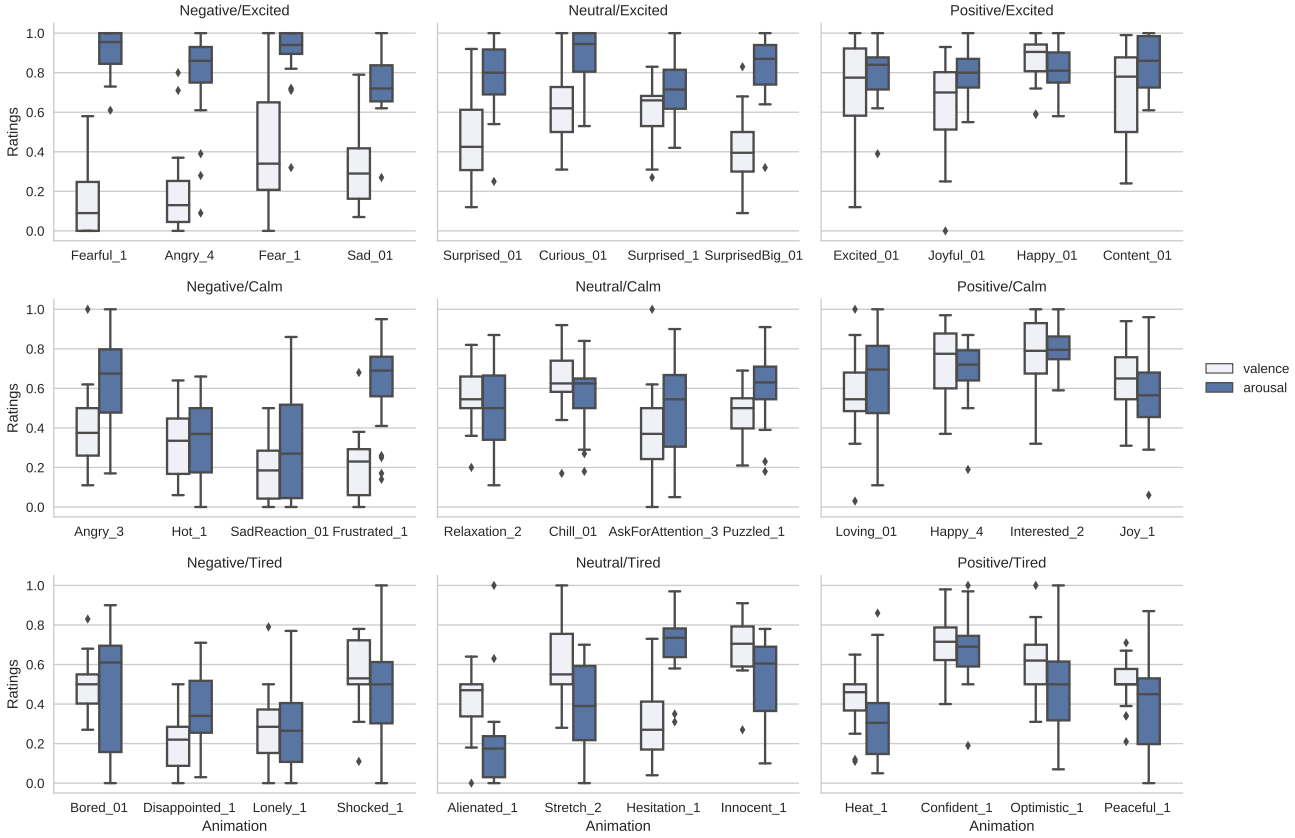

Fig. 4.4 Boxplots of the ratings collected from the whole group of participants ($N = 20$) per preassigned class of affect (9 classes with 4 animations each). The classes are labelled as combinations of valence and arousal levels. Individual animations are referenced with the original categorical tag assigned by the animators during the creation phase.

Fig. 4.5 shows all the ratings plotted in the valence-arousal space. The sparsity of data points in the low arousal subspace is more pronounced in the high valence area. Moreover, an accumulation of ratings can be observed in the area of neutral valence (valence = 0.5). More specifically, 98 ratings out of 720 in total are accumulated at valence equal to 0.5. This is not the case for arousal (only 33 ratings out of 720 were found at 0.5). For both variables equal to 0.5 (origin of the space), only 10 out of 720 ratings were found. To better understand this effect, we plotted the kernel density estimates (KDE) for the ratings concerning the five different levels of raters' confidence in Fig. 4.6.

The KDEs in the valence graph in Fig. 4.6 show a peak at 0.5 (neutral valence) for the ratings submitted with low confidence, i.e., from 1 to 3 ("not confident at all" to "somewhat confident"). This trend appears stronger for male raters in confidence levels 1 and 3, and

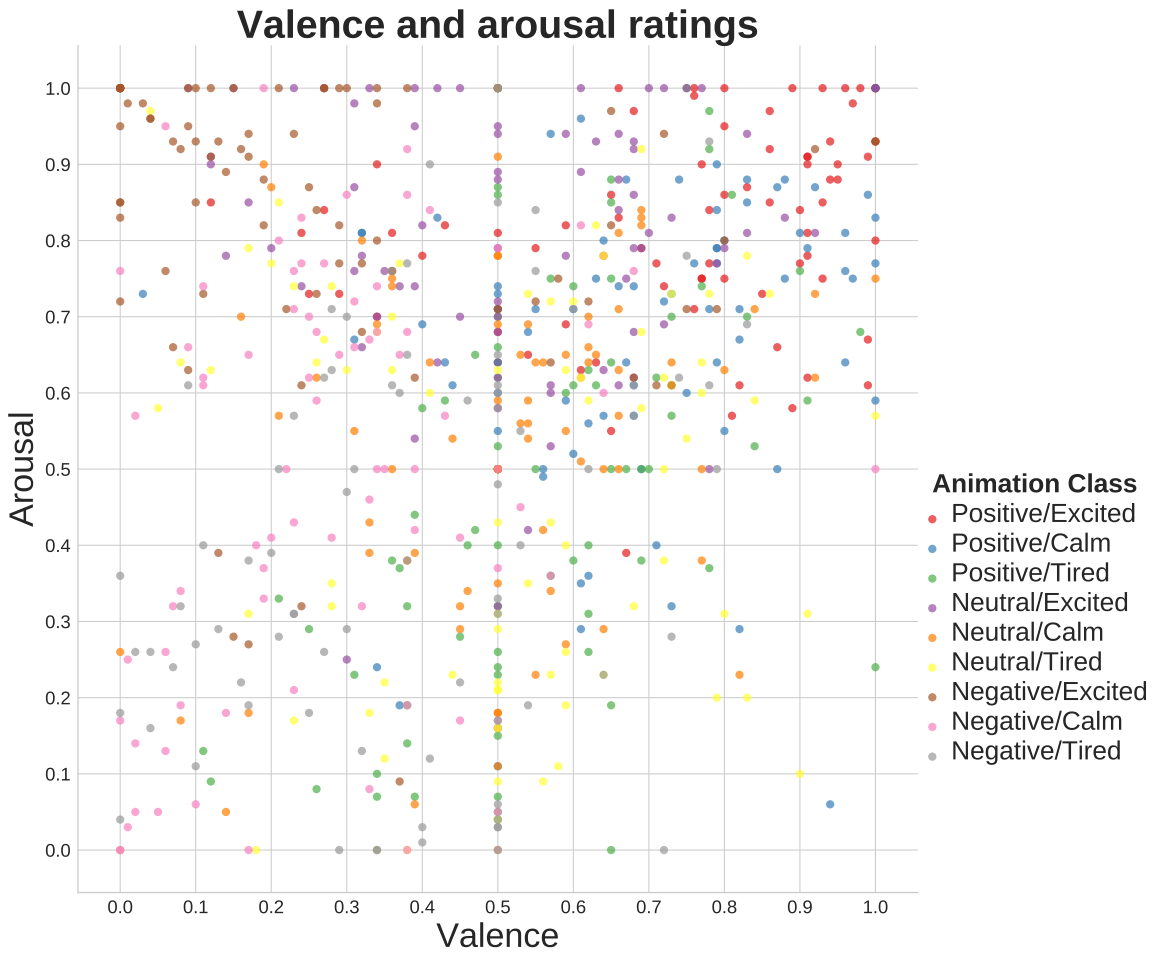

Fig. 4.5 Valence and arousal ratings from all animations and raters colour-coded according to the pre-assigned animation class.

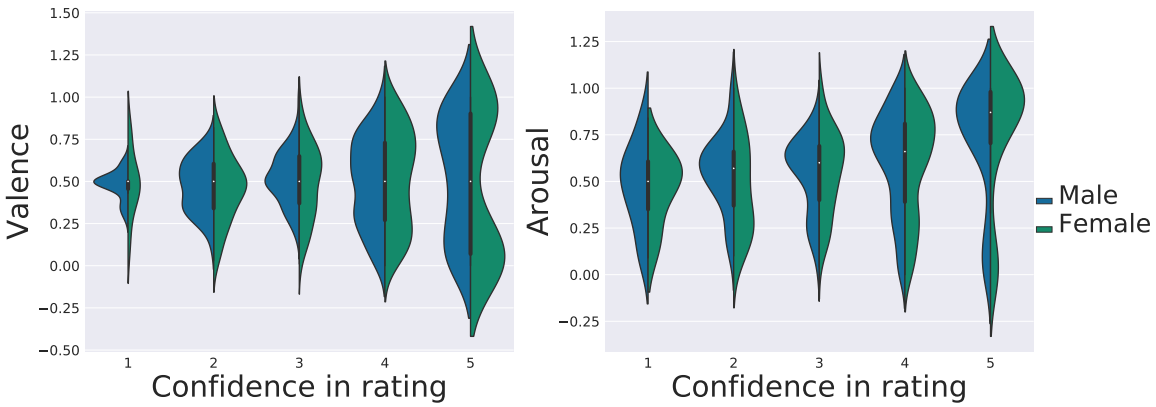

Fig. 4.6 Kernel density estimates of the ratings with respect to the raters' confidence level (0: "not confident at all" to 5: "very confident").

female raters in confidence level 2. This trends could mean that when participants are not confident in their interpretation, they tend to resort in a neutral appraisal of valence.

The peak at 0.5 is not as prominent in the KDE graphs of arousal (Fig. 4.6), at least not above the lowest level of confidence. Furthermore, the arousal KDEs appear to be more negatively skewed than valence KDEs for all confidence levels. This trend might imply that the polarization we noticed before, with the sparse ratings in the low arousal subspace, is not related to difficulties in interpreting the animations (lack of confidence in judgement). Instead, it supports the possibility that participants indeed perceived fewer animations as of low arousal. Moreover, for both dimensions, the more extreme the ratings are, the higher the confidence in the judgement, both for male and female participants. Finally, we examined for statistical differences in raters' confidence levels between valence and arousal. A one-sided Wilcoxon Signed Rank Test showed that raters' confidence in the arousal ratings was significantly higher than the confidence in the valence ratings ($z = -1.90$, $p = .03$). This result could indicate that valence is harder to estimate than arousal.

4.3.4 Intra-rater reliability

To test intra-rater consistency, we repeated the presentation of the first three animations at the end of the main session, and we measured the intra-rater mean squared error (MSE) and the intraclass correlation.

Intra-rater reliability results

The average MSE across 20 participants was 5.26% ($SD = 5.96$, $50th$ percentile $= 2.87$, $95th$ percentile $= 14.13$) for arousal, and 3.97% ($SD = 4.42$, $50th$ percentile $= 2.91$, $95th$ percentile $= 10.24$) for valence. Hence, MSE was less than 25% in all cases and less than 15% in 95% of the cases.

The intraclass correlation coefficient (ICC) for intra-rater reliability was estimated as a 2-way mixed-effects model with the definition of *absolute agreement* since we are testing a paired group. The formula for the estimation is the following:

$$ICC(2,k) = \frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}} \tag{4.1}$$

where k denotes the number of repeated samples, MS_R is the mean square for rows, MS_E is the mean square for error, MS_C is the mean square for columns, and n is the number of items tested [126].

The result was $ICC = 0.77$ with 95% confidence interval $0.67 < ICC < 0.84$, and $F(119,119) = 4.41$, which is interpreted as a moderate to good reliability. We discuss more details about the interpretation of the ICC coefficients in the next section.

4.3.5 Inter-rater reliability

To assess the inter-rater reliability, we used intraclass correlation, one of the most widely-used statistics for assessing ratings reliability [206, 161], applicable to ratings of interval type. For fully crossed design, i.e., each item (animation) was rated by all the participants, ICC can provide a good estimation of the reliability with adequate control of systematic bias between the raters [92].

The selection of the appropriate ICC form depends on three parameters; the model, the type and the definition of ICC [126]. The model we have chosen is the two-way random effects model since we randomly selected our raters from a larger population, and we want to be able to generalize the results to a larger population with similar characteristics. The type of ICC is the mean of the ratings since we will use the aggregated ratings. Finally, the definition of the ICC form is that of *consistency* (instead of *agreement*) since we are not interested in the absolute agreement between the raters, but on how well do the ratings correlate in an additive manner. The formula is given in Eq. 4.2 and is defined for consistency instead of agreement:

$$ICC(3,k) = \frac{MS_R - MS_E}{MS_R} \tag{4.2}$$

where k denotes the number of the raters, MS_R is the mean square for rows and MS_E is the mean square for error [126].

The ICC estimates we present in the results section were calculated using the function *icc* (parameters: "twoway", "consistency", "average") from the *irr* package [80] of R [186]. Finally, for the interpretation of the results we are following Koo et al. [126]: poor for values < 0.5, moderate for values between 0.5 and 0.75, good for values between 0.75 and 0.9, and excellent for values > 0.9. The evaluation thresholds are applied to the 95% confidence intervals and not the ICC value itself.

Inter-rater reliability results

The resulting ICCs (Table 4.4) for the whole group of the participants are in a high range for both valence and arousal indicating high consistency and that raters perceived the core-affect expressed by the animations similarly. We also estimated the ICC for males and females, as well as with respect to the confidence in participants' judgements, by taking the ratings of

Table 4.4 Intraclass correlation coefficients of inter-rater reliability

		Intraclass Correlation	95% Confidence Interval		F Test With True Value 0		
			Lower Bound	Upper Bound	Value	df1	df2
Valence	Group	0.95	0.93	0.97	21.2	35	665
	Female	0.91	0.86	0.95	11.1	35	280
	Male	0.92	0.87	0.95	12	35	350
	Low Conf.	0.81	0.58	0.94	5.19	9	171
	High Conf.	0.98	0.95	0.99	47	9	171
Arousal	Group	0.95	0.92	0.97	18.7	35	665
	Female	0.91	0.85	0.95	10.7	35	280
	Male	0.89	0.83	0.94	8.96	35	350
	Low Conf.	0.85	0.67	0.96	6.68	9	171
	High Conf.	0.97	0.94	0.99	35.3	9	171

the 10 highest in confidence and 10 lowest in confidence animations. In the case of the lower confidence ratings, ICC is lower, indicating lower consistency among raters compared to that of higher confidence, but still in an acceptable range.

The Feldt test [72, 128] for statistical inference between the ICC coefficients for female and male participants did not reveal any statistical differences (valence: $F(1,18) = 0.03$, $p = .88$, arousal: $F(1,18) = 0.08$, $p = .75$). Applying the same test for paired samples revealed significant differences between ICC coefficients for low and high confidence in the ratings (valence: $t(18) = 17.26$, $p < .001$, arousal: $t(18) = 9.38$, $p = .001$). The results for the Feldt test were obtained with R [186], and the *cocron* package [63].

4.3.6 Final labels of valence and arousal

Supported from the inter-rater reliability results, we derived the final set of the aggregated annotations by taking the mean across raters for each animation. The final annotations are plotted in the valence-arousal space in Fig 4.7, along with the original tags of the animations (assigned by the animators during the creation phase) and colour-coded according to the preassigned class. Detailed statistics and raters' confidence levels for the final annotation set can be found in Appendix A, Table A.2. A video depicting the animations and the derived labels of valence and arousal is available online[4].

[4]Animation set video: https://youtu.be/Zo-K0uczhKc

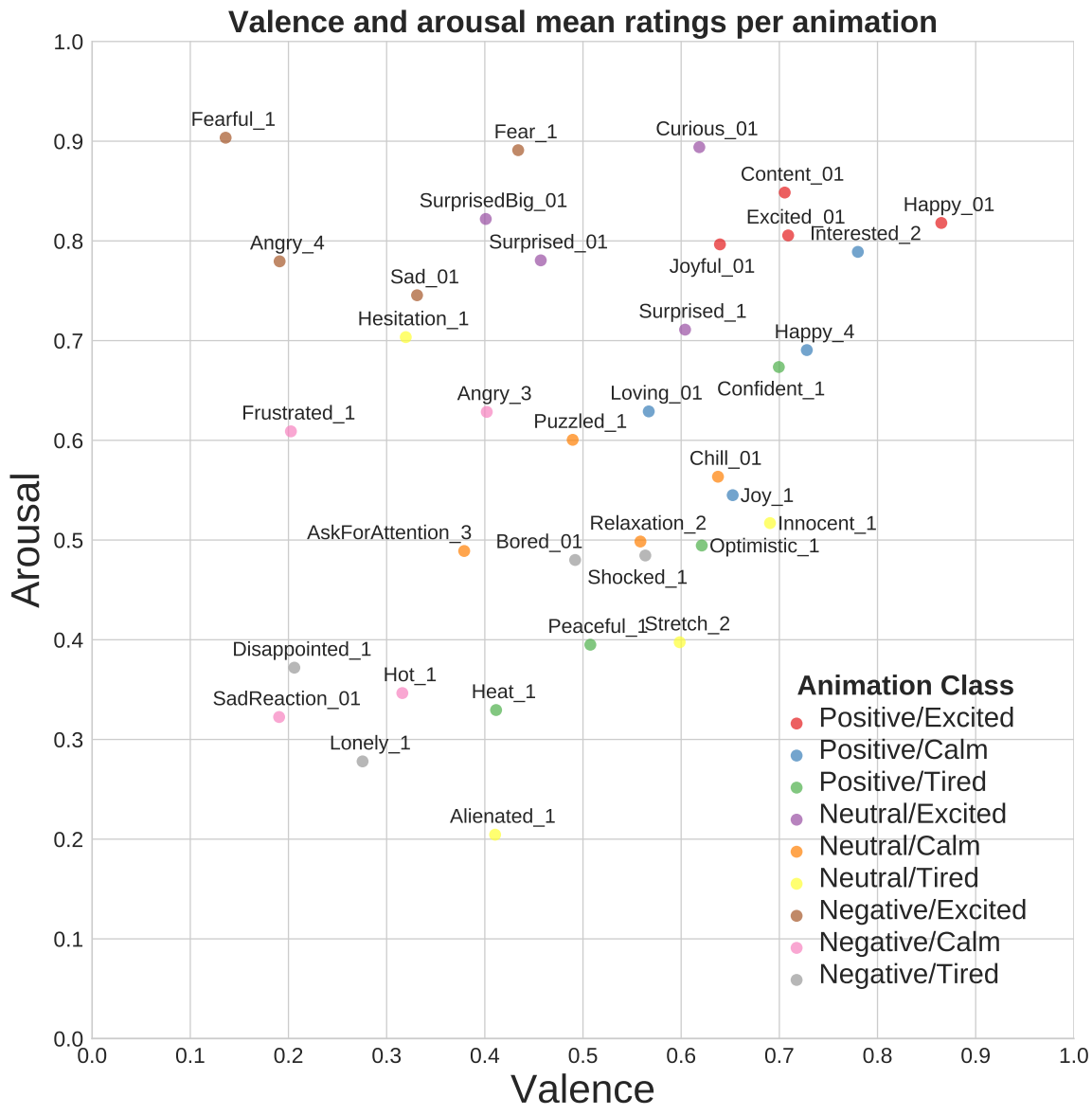

Fig. 4.7 The final set of valence and arousal labels derived from the means across raters. They are plotted with the original categorical tags of the animations (given by the animators in the creation phase) and colour-coded according to our pre-assigned classes.

4.4 Discussion and conclusion

The work presented in this chapter used a set of pre-designed, nonverbal animations created with the pose-to-pose method for the Pepper robot to simulate EBL. Our objective was to annotate the animations with continuous interval, dimensional affect labels of valence and arousal. To validate the collected ratings we tested the inter-rater reliability with the intraclass correlation (applied across subjects per animation) which was found in an excellent range (ICC > 0.90) for both valence and arousal. This result indicates very high consistency among raters and shows that each animation is perceived similarly by the participants.

The intra-rater MSE was below 23.8% for all participants and below 14.13% for 95% of the participants. Intra-rater ICC was found moderate to good, indicating an acceptable within-rater consistency. The final set of aggregated labels for each animation was derived from the mean across all participants. The means are plotted in Fig. 4.7, and the complete descriptive statistics are given in Appendix A, Table A.2.

We also aimed to explore trends in the ratings. We found that when the level of confidence in the judgment was low for valence, the participants tended to evaluate the expression as neutral. We then examined statistical differences in raters' confidence level between valence and arousal and found that the latter was significantly higher, which might indicate that valence was harder to rate. This observation appears in agreement with other evidence indicating that arousal is more easily perceived from bodily expressions than valence [123], while facial expression appears as a more stable visual modality for valence recognition [88]. Furthermore, previous work suggests that context has a significant influence on how humans perceive valence [14], and in the present experiment the animations were presented in a context-free setup. Under the light of this evidence, the fact that the animations were displayed: 1) on a robot with constrained facial expression, and 2) in a context-free setup, could explain why participants find it harder to evaluate valence.

For the arousal ratings, we observed that the lower arousal subspace was more sparsely covered compared to the medium- and high-arousal subspaces, especially in the higher-valence subspace (*Positive/Tired*). This trend could be related to the subjective pre-assignment of the animations in the nine classes; it might be the case that the pre-assignment did not succeed to cover the affect space fully. Another possibility could be that the perception of arousal differs between the group who conducted the pre-assignment (two researchers and an animator highly accustomed to the robot) and the participants who had minimal or no experience with Pepper. In the second case, a question that arises is whether the novelty effect contributes to inflated arousal ratings. Finally, a different explanation could be that this trend might indicate the challenges related to designing animations that map to this subspace of low arousal and high valence.

4.4.1 Limitations

With regards to the limitations of this work, the choice of the two-dimensional model of emotion might be considered as too limited since it has been shown that subtle variations between certain emotions that share common core affect, e.g., fear and anger, might not be captured in less than four dimensions [74]. Nevertheless, core affect is still considered valid as a low dimensional structure representing aspects of the full-blown emotional space [196]. In our future work, we would like to explore how additional dimensions, like dominance and unpredictability, contribute to enabling user discrimination of emotional expressions.

Another limitation is that we did not systematize and balance the animation set in terms of the interactive modalities; motion, eye LEDs and sound. Nevertheless, with this experiment, it was not our intention to examine how different combinations of modalities impact emotion perception. This is a very interesting question, but our present work treated each animation in a holistic manner.

Finally, the present study only involved one robotic platform, the Pepper robot, and consequently, the observations might not transfer to other robots with a different design, level of embodiment or expressive capabilities. Emotion expression with body language is inherently dependent on the embodiment level, and the design, especially the more complex the expressions are, and the particular characteristics of Pepper might have evoked impressions that would be different in another platform. This is a native challenge in studying the perception of emotion expression in HRI, and comparison between different approaches appears challenging as we discussed in Chapter 3. Moreover, the animations we used are designed for the particular platform by professional animators, and their structure cannot be translated and adapted for other robots automatically. Although our work can not directly generalize to other platforms, hopefully, it could potentially serve as a case study from which other researchers could shape hypotheses to test with different robots.

4.4.2 Next steps

With the completion of our first study, two different future directions emerged. The first direction, continues on the same path of the current study, exploring how people perceive robotic EBL. A dominant concern would be how the perception is affected by contextual stimulus alongside the robotic expressions. Would valence and arousal ratings differ significantly when participants are presented with the same animations as a response to congruent or incongruent contextual information? Another research question in this path is the perception of dominance and how it correlates with valence and arousal.

Although we find these research questions essential for a deeper understanding of how humans perceive robotic EBL, our main interest lies in the second direction, which concerns the EBL synthesis. The next steps in our project focus on deep learning methods to accomplish robotic EBL synthesis. More specifically, we decided to use the Variational Autoencoder (VAE) framework [120], a powerful generative model. In the next chapter, we will provide a detailed description of the framework, and in Chapter 6 we will present our first applied study in which we train the VAE with the animation set we labelled in the present study.

Chapter 5

Generative modelling and the variational autoencoder framework

5.1 Introduction

Deep learning algorithms can be broadly divided into two approaches: discriminative modelling and generative modelling. The former has been the driving force behind most advances in basic and applied deep learning research during the last two decades. Essentially, it is synonymous to supervised learning classification task, where a deep network learns a function that maps an input to a label, or in probabilistic terms, the model learns to estimate the probability of a label y given an observation $\boldsymbol{x}$. A significant breakthrough in discriminative modelling occurred in the 1990s with convolutional neural networks (CNNs) [134] for image classification. The successful performance of CNNs, especially after the AlexNet [129] surpassed human performance in image classification in the ImageNet challenge in 2012, paved the way for the dramatic increase of the interest we are currently witnessing in deep neural networks.

On the other hand, the theoretical breakthroughs that allowed for deep generative modelling applications occurred more recently, with Variational Autoencoders (VAE) [120] in 2013, Generative Adversarial Networks (GAN) [85], and autoregressive models such as PixelRNN [216] in 2016. The area has attracted significant attention since, with a lot of theoretical and applied research dedicated to advancing the field. In generative modelling, the goal is to learn the underlying distribution of the observed data, i.e., to learn the probability of observing $\boldsymbol{x}$, without the necessity of labels y. Learning this probability allows us to sample the distribution and generate new realistic synthetic samples similar to the original data. This is very useful in applications where many simulations are necessary to train for

instance a reinforcement learning agent, or in creative applications, such image synthesis, image denoising, image inpainting, image super-resolution, prediction of future frames in video, text and speech generation, on-demand generated art.

The rapidly increasing interest in generative modelling is also fuelled by the current availability of a huge amount of unlabelled data which cannot be used in supervised learning. Furthermore, the advances in generative models bring new hope for a more sophisticated form of artificial intelligence, in which machines perceive the underlying essence of the data, instead of just using their statistical properties to classify them. This is a core concept in *representation learning*, [23], where the aim is to identify and disentangle the hidden explanatory factors of the data, and thus obtain compact and informative data representations (latent representations) which can subsequently be used in downstream tasks.

The three generative model approaches have different advantages and limitations. Briefly, the VAEs are very stable during training, straightforward to evaluate (with log-likelihood), but they tend to generate samples that discard high-frequency details (e.g., blurry images). On the other hand, GANs generate very sharp samples, but there is no exact way to evaluate them except by visualising the generated samples. Furthermore, they are notorious for their unstable training dynamics, and they require many training data. Autoregressive models have stable training dynamics and can be evaluated with log-likelihood. However, they are slower to train because each sample is fed element by element to the network, and they do not learn directly latent representations of the data that could be useful in representation learning.

In this work, we have chosen the VAE model to generate robotic emotional body language. The VAE has been a very influential approach. It is a very efficient and elegant probabilistic model that can be trained to code the data into latent representations, which are more compact, informative and possibly disentangled. In fact, in terms of learning informative and disentangled latent representations, the VAE shows robust results [234]. Although our primary goal is to generate realistic samples, the representation learning perspective is a desirable property for our future work.

Furthermore, the VAE model is more stable to train with a small training set of examples, and since we do not generate images but essentially body postures, our results are not affected by the lack of sharpness. One could argue that a sequential model from the autoregressive paradigm could better fit our goals since body language animations are essentially sequences of postures. However, although the VAE input is not processed sequentially (we will see that we train with shuffled postures), the latent space is constrained so that we can sample and interpolate the samples producing a very smooth sequence of postures.

The rest of this chapter can be summarised as follows: we begin with a description of standard and regularised autoencoders which are simpler non-generative models whose

similarities and differences with the VAEs can pave the way for a deeper understanding of the more complicated VAE model and its great value. Next, we present the VAE framework from the probabilistic graphical model point of view, the derivation of its learning objective, and a smart modification that allows optimising the objective with stochastic gradient descent. We finish this chapter with the deep learning perspective of the VAE model and some related challenges and solutions.

5.2 Autoencoders

We will begin with a brief discussion of the standard autoencoder, a type of artificial neural network which preceded the VAE and it is used for dimensionality reduction. The autoencoder is not a generative model, but since there are some similarities with the VAE in the structure, it can be useful to understand its more straightforward theory and then proceed with the VAE theory. It is not easy to track the first appearance of the autoencoders in literature due to the gradual formation of the concept, which was often expressed with different terminology (for instance auto-associative multilayer perceptrons [32, 8]), before advancing at the current state [99].

An autoencoder consists of two sub-networks, the encoder and the decoder, which can be implemented with any class of feedforward neural networks. It is trained holistically with the usual techniques that apply to feedforward models, i.e., gradient descent with backpropagation. The training set consists of sample vectors $\boldsymbol{x}$, each a collection of one or more variables of interest, either continuous or discrete. Multi-dimensional variables are flattened first and then concatenated with the rest. The input vector $\boldsymbol{x}$ passes through the encoder which consists of one or more layers of progressively smaller dimensionality; thus the output, which is called *encoding* $\boldsymbol{z}$, is a compact representation of lower dimensionality compared to the input $\boldsymbol{x}$. Subsequently, the encoding $\boldsymbol{z}$ passes through the decoder network layers, which are progressively wider, until a final layer with the size of the original input. The final layer thus reconstructs the dimensionality of $\boldsymbol{x}$ and outputs $\tilde{\boldsymbol{x}}$. To learn an effective encoding of the data, the whole network is trained with a loss function (Mean Square Error or cross-entropy) that penalises the difference between the original input $\boldsymbol{x}$ and the reconstructed output $\tilde{\boldsymbol{x}}$. Technically, the autoencoder learns to copy the input to the output. However, the bottleneck (the low dimensional middle layer between the encoder and the decoder) learns a compressed representation as a byproduct of the main copy task.

The autoencoder works well as a dimensionality reduction technique, and it can be used for extracting features to train other networks. As a matter of fact, an autoencoder with linear activation functions would learn the same features as the Principal Component Analysis

(PCA) algorithm, the oldest and most widely used method for discovering the maximal variance directions in high-dimensional data [97]. Nevertheless, an autoencoder network with nonlinear activation functions is more powerful compared to PCA.

5.2.1 Regularized autoencoders

Although autoencoders perform well in the compression task, they are limited in two important aspects. First, the compressed latent representation is not necessarily guaranteed to preserve all the essential information in the original input. The network may be learning sub-optimal encodings of the input, which do not represent the underlying factors of variation in the data, but they allow a good reconstruction. In other words, the model is prone to overfitting the training set instead of learning useful generalisations. This happens because there is no constraint imposed on the learned representation, and the network only trains on how to produce a good reconstruction. A second limitation arising from this is that the autoencoder cannot be used as a generative model. Since the latent space is not constrained, sampling from it would lead to noisy output whenever we sample from a subspace that was not included in the training.

There have been several architectures proposed to tackle these sort of limitations by applying different terms of regularisation in the loss functions. Essentially, the loss function is modified to involve additional constraints such as robustness to noise or masked inputs, the sparsity of the latent layer, or the smallness of the derivative of the representation. These constraints aim in incentivising the model to learn more useful representations of the input data than just salient features that allow a good reconstruction. We will briefly present three models using regularisation to improve the overfitting problem of the autoencoder. For a more detailed account of these models, we refer the interested readers to [84] and the citations provided in the following discussion.

Denoising autoencoders [218] are fed with a corrupted version of $\boldsymbol{x}$ (noise is added or some dimensions are masked with zeros), but are trained to recover the original input. The loss function is defined as in the standard autoencoder, penalising the difference between the original and the reconstructed, but the network only encodes and decodes the corrupted version of the original data. This design tackles to some extent the problem of overfitting on particular dimensions of the input data distribution and can also be seen as a kind of dropout since a fixed proportion of randomly chosen dimensions of the input are replaced with zeros.

Sparse autoencoders have an additional regularisation term in their loss function, the sparsity penalty, which constrains the latent representation and prevents overfitting. This is achieved by obliging each latent unit to be inactive most of the time, and hence only a subset of the latent units are active at each time step, turning the latent representation into a sparse

one. In k-Sparse Autoencoder [148] the sparsity penalty is implemented so that only k units with the highest activation values are kept, while the rest are zeroed out.

Contractive autoencoders [189] apply a constraint on the latent representation by penalising its sensitivity to the input perturbations. Small changes in the input should not be encoded far apart from each other in the latent space. That is accomplished by an extra term on the loss function, besides the reconstruction error, which penalises large derivatives of the encoder's output. Intuitively, this means that the model is encouraged to learn a latent space in which similar inputs have similar encodings, or, in other words, a neighbourhood of input datapoints is *contracted* to a smaller output neighbourhood. Although this contraction effect permits the formation of some structure in the latent space (instead of randomly scattered encodings in the standard autoencoder), this structure is local, that is, only the perturbations of an input datapoint are encoded in a contracted neighbourhood, while two different inputs can be very far apart from each other.

Although these three models enhance the autoencoder architecture with the ability to learn more useful features instead of just being an efficient copy machine, additional constraints must be imposed to be able to obtain a generative model. This brings us to the Variational Autoencoder framework.

5.3 Variational autoencoders

The goal of the Variational Autoencoder (VAE) [120, 110, 121], architecture is to learn a probability distribution $p^*(\boldsymbol{x})$ that represents the unknown underlying process that generates the observed data. Subsequently, this distribution can be sampled to synthesise purely novel content, that is, new, previously unseen data similar to the original. This cannot be accomplished with a standard autoencoder, a deterministic model that always encodes and reconstructs its input in an identical way, essentially performing a memorisation task. Instead, the VAE achieves the reconstruction task by way of a stochastic process deeply rooted in the theory of Bayesian networks.

5.3.1 The probabilistic graphical model representation

Bayesian networks define a class of probabilistic graphical models, a branch of machine learning that uses probability distributions to model real-world events and make useful predictions about them. This is a powerful and elegant approach in problem-solving since probability distributions capture the inherent uncertainty involved in many real-world phenomena. Bayesian networks can effectively describe a set of variables and their conditional

dependencies with directed acyclic graphs (DAGs). This graphical representation is a natural and effective way to highlight causality relationships. In DAGs, the vertices correspond to the problem space variables, and the directed edges represent the various dependencies among them. The graph is acyclic because there are not any edges directed from a child node towards a parent node. Next, we will use the probabilistic graphical model representation to describe the two main processes of the VAE framework: the generative and the inference process.

5.3.2 The generative process

The generative process of the VAE can be represented with a directed latent-variable probabilistic graphical model. This is essentially a probabilistic DAG that involves latent variables, typically denoted by $\boldsymbol{z}$, in addition to the observed ones, denoted by $\boldsymbol{x}$. Latent variables are unobserved variables that are not included in the dataset explicitly; instead, they are hidden in the data, and they have to be inferred by the model. This graphical model representation is shown in the schematic of Fig. 5.1, where N is the number of the observed datapoints $\boldsymbol{x}$ and each of them depends to a local latent $\boldsymbol{z}$. The node θ denotes the parameters of the model which are going to be learned. It lies outside the plate notation because θ is global to the model, that is, all $\boldsymbol{x}$ datapoints depend on it as the direction of the edge implies. Intuitively, the latent variables $\boldsymbol{z}$ contain information about the hidden structure of $p^*(\boldsymbol{x})$, the actual probability distribution that generates the observed data, and the parameters θ weigh this information to generate samples successfully.

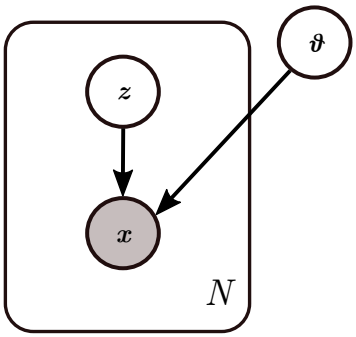

Fig. 5.1 The generative process $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ of the VAE as a directed probabilistic graphical model.

More concretely, the graphical model expresses a joint distribution whose factorization describes the generative process:

$$p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{z}) p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) \tag{5.1}$$

The distribution $p_\theta(\boldsymbol{z})$ is called the *prior* distribution since it does not depend to any other variable (unconditional). According to Eq. 5.1, we first sample a latent datapoint $\boldsymbol{z}$ from $p_\theta(\boldsymbol{z})$, and then we use it to condition the generation of $\boldsymbol{x}$. The conditional distribution $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ is the *likelihood*.

5.3.3 The inference process

The above formulation describes the generative process represented by the probabilistic graphical model, but an inference process is also necessary to derive optimal values of the latent variables. The inference process is depicted as a graphical model in Fig. 5.2). The inference process aims to ensure that the latent space of the model encompasses all the meaningful information that will enable the generation of faithful samples. In other words, we want to obtain latent variables $\boldsymbol{z}$ that maximise the likelihood of generating the $\boldsymbol{x}$ samples in our training set.

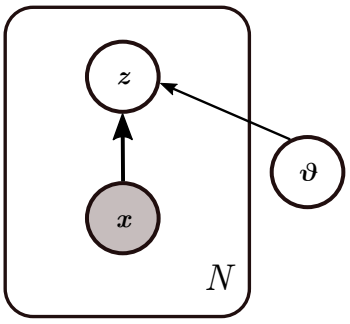

Fig. 5.2 The inference process $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ of the VAE as a directed probabilistic graphical model.

The inference process can be achieved by calculating a *posterior* distribution $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$, which can be written according to the Bayes rule as follows:

$$p_\theta(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p_\theta(\boldsymbol{z}) p_\theta(\boldsymbol{x} \mid \boldsymbol{z})}{p_\theta(\boldsymbol{x})}. \tag{5.2}$$

There are two requirements for solving this equation: 1) to define the prior distribution $p_\theta(\boldsymbol{z})$, in the sense of deciding what kind of hidden attributes of the data should be captured by the latent variables $\boldsymbol{z}$, and 2) to compute the quantity in the denominator, $p_\theta(\boldsymbol{x})$, which is called the *model evidence* or *marginal likelihood* of the data. The first requirement is straightforward, and there is no need to explicitly define what information should each dimension of $\boldsymbol{z}$ capture, or describe any dependencies between the $\boldsymbol{z}$ dimensions [64]. This is because the VAE framework avoids strong prior assumptions, and instead, it draws samples

of $\boldsymbol{z}$ from a simple distribution. The usual choice for the prior $p_\theta(\boldsymbol{z})$ is a centered isotropic multivariate Gaussian:

$$p_\theta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \boldsymbol{I}), \tag{5.3}$$

where the mean is a vector of zeroes, the covariance matrix is the identity matrix $\boldsymbol{I}$, and since the distribution is fixed, θ is empty, i.e., the prior lacks parameters. We will see later that the implementation of the generative process with a multilayer neural network, which is a powerful function approximator, is efficient for mapping the normally distributed $\boldsymbol{z}$ to informative latent structure that allows for faithful reconstruction.

The second requirement, computing the model evidence $p_\theta(\boldsymbol{x})$ in the denominator of Eq. 5.2, is more complicated. By using the law of total probability, the evidence can be written in relation to $\boldsymbol{z}$ as follows:

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{z}) p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) d\boldsymbol{z}. \tag{5.4}$$

Evaluating 5.4 requires to integrate over a very large number of $\boldsymbol{z}$, and since there is no analytic solution to do this, the computation of the posterior is considered intractable (it cannot be evaluated or differentiated). The VAE framework addresses this issue by using the *Auto-Encoding Variational Bayes (AEVB)* algorithm [120] to approximate the posterior distribution $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$.

5.3.4 The learning objective

AEVB emerges from *variational inference* methods [28], in which an inference problem is approached as an optimization problem, and instead of estimating a complex probability distribution, such as the true posterior $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$, we approximate it with a simpler, known probability density which is tractable. Note at this point that another prevalent method for approximating posterior densities is the Markov Chain Monte Carlo (MCMC) sampling, but compared to variational inference, MCMC tends to be less fast and easy to scale to large data.

In variational inference, we first define a family of tractable distributions Q over the latent variables $\boldsymbol{z}$. Then we optimize for finding a family member $q \in Q$ (more accurately, finding the parameters of the q distribution, the *variational parameters*), that is the most similar to the true posterior. This approximate posterior is introduced in the VAE framework with a parametric inference model $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$, where ϕ denotes the variational parameters which we need to optimize so that $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \approx p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$. In figure 5.3 we depict both the inference and the generative process in one probabilistic graphical model.

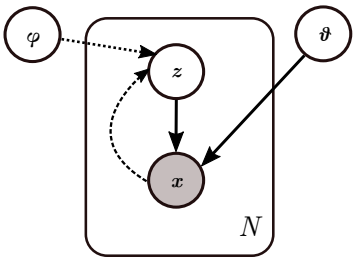

Fig. 5.3 The full VAE model as a directed probabilistic graphical model. The inference process flow is shown with dashed lines, while the generative process with solid lines.

The similarity between the approximate and the true posterior distribution is measured with the Kullback-Leibler divergence D_{KL}, a metric from the field of information theory, which estimates the disparity between two distributions [113]. The measure is asymmetric, i.e., $D_{KL}(q \parallel p) \neq D_{KL}(p \parallel q)$, non-negative, and it equals zero when the distributions are the same. Therefore, for the optimization problem of finding a good q approximation, KL divergence needs to be minimized, i.e., the lower the KL divergence, the higher the similarity between the two distributions:

$$q_{\phi}^{*}(\boldsymbol{z} \mid \boldsymbol{x}) = \underset{q(z|x) \in Q}{\arg\min} D_{KL}(q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_{\theta}(\boldsymbol{z} \mid \boldsymbol{x})). \tag{5.5}$$

Finding q^{*} ensures finding the best approximation within the Q family of distributions. The degree of complexity characterizing the selected Q family determines the complexity of the optimization problem. Theoretically, any family can be used, but in practice, a common choice to simplify the optimization is to use the multivariate Gaussian with mean $\boldsymbol{\mu}$ and a diagonal covariance matrix, that is, a scalar variance multiplied by an identity matrix:

$$q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^{2}\boldsymbol{I}). \tag{5.6}$$

The D_{KL} between the approximate posterior and the true posterior is given by the following formula:

$$D_{KL}(q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_{\theta}(\boldsymbol{z} \mid \boldsymbol{x})) = \mathbb{E}_{\boldsymbol{z} \sim q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}[\log q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x}) - \log p_{\theta}(\boldsymbol{z} \mid \boldsymbol{x})], \tag{5.7}$$

where the expectation is taken with respect to all $\boldsymbol{z}$ drawn from $q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x})$. Note that in the expectation of the right-hand side, $\boldsymbol{z}$ denotes instances of the random variable, while $\boldsymbol{z}$ in D_{KL} of the left-hand side denotes the random variable generally. The expectation subscript $\boldsymbol{z} \sim q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x})$ is removed from the following equations to avoid clutter. In the next step, Bayes' rule is applied on the true posterior $p_{\theta}(\boldsymbol{z} \mid \boldsymbol{x})$ of the right-hand side:

$$D_{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) = \mathbb{E}[\log q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) - \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) - \log p_\theta(\boldsymbol{z})] + \log p_\theta(\boldsymbol{x}), \quad (5.8)$$

where the term $\log p_\theta(\boldsymbol{x})$ is taken out of the expectation because it does not depend on $\boldsymbol{z}$ over which the expectation is taken. Next, by negating both sides, rearranging the terms and using the property of linearity of expectation, we obtain the following:

$$\log p_\theta(\boldsymbol{x}) - D_{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) = \mathbb{E}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - E[\log q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) - \log p_\theta(\boldsymbol{z})], \quad (5.9)$$

where the second expectation on the right-hand side is equal to D_{KL} between the approximate posterior and the prior, and thus, can be rewritten as follows:

$$\log p_\theta(\boldsymbol{x}) - D_{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) = \mathbb{E}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})]. \quad (5.10)$$

At this point, we notice that the intractable marginal likelihood (the evidence) is once again involved in $\log p_\theta(\boldsymbol{x})$. However, this time, by rearranging the terms, we can derive a *variational lower bound* for the marginal log-likelihood, also called the *evidence lower bound* (ELBO), which is tractable and therefore, it can be optimized via stochastic gradient descent:

$$\log p_\theta(\boldsymbol{x}) = \underbrace{\mathbb{E}_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})]}_{ELBO} + D_{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{x})). \quad (5.11)$$

As we mentioned before, Kullback-Leibler divergence is always non-negative, therefore, the last term on the right-hand side, $D_{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{x}))$ will always be greater or equal to zero. This property turns the ELBO part of the equation into a lower bound for the marginal log-likelihood $\log p_\theta(\boldsymbol{x})$:

$$\begin{aligned} \log p_\theta(\boldsymbol{x}) &\geq ELBO \\ &\geq \mathbb{E}_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})]. \end{aligned} \quad (5.12)$$

Hence, maximising the ELBO with respect to the parameters θ and ϕ accomplishes our optimisation goal by enforcing two things simultaneously. First, the marginal log-likelihood

$\log p_\theta(\boldsymbol{x})$ is maximised, which means that we obtain a better fit of the data, or equivalently, the model stands better chances to generate the data. Second, we minimise the KL divergence of the approximate posterior from the true posterior, thus we approximate $q*$ (this becomes clear when we move the ELBO term on the left-hand side of the Eq. 5.11). There is another way to derive the ELBO objective [112] through Jensen's inequality, but we preferred the above derivation presented in [120, 64], as more intuitive about the VAE framework components.

Finally, by negating the ELBO objective we derive a loss function which can be minimized to learn the parameters θ and ϕ:

$$\mathscr{L}_{\theta,\phi}(\boldsymbol{x}) = \underbrace{-\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})]}_{\textit{reconstruction error}} + \underbrace{D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})]}_{\textit{regularization term}}. \tag{5.13}$$

In the loss function equation, the *reconstruction error* is the negative expected log-likelihood, and it encourages latent $\boldsymbol{z}$ values that the generative model can use to reconstruct faithfully the given $\boldsymbol{x}$. On its own, this term is, in fact, equivalent to the standard autoencoder loss function, which as we discussed before is, in essence, a copy machine. It is the *regularization term* in 5.13 (also called variational loss), that makes the VAE a powerful generative model that learns useful data representations. The regularisation term in the VAE objective encourages the inference model to learn $\boldsymbol{z}$ values that are close to the prior distribution, and this way it applies a constraint on the latent variables' space, resulting in a continuous manifold from which we can sample and generate new realistic samples.

5.3.5 Optimization with stochastic gradient descent

So far, we discussed how the AEVB algorithm is using variational inference and the evidence lower bound (ELBO) to deal with the intractable posterior. Furthermore, we presented how the learning objective is defined for the joint optimisation of the inference and generative components of the VAE. Next, we will see how this objective can be optimised (that is, how to maximise the ELBO or equivalently minimise the loss function 5.13), with a gradient estimator. The optimisation entails some challenges, and we will see how the AEVB algorithm provides a very elegant solution to address them.

Optimizing the loss function of Eq. 5.13 requires taking its gradient with respect to the generative model parameters θ and the variational parameters ϕ of the inference model:

$$\nabla_{\theta,\phi}\mathscr{L}_{\theta,\phi}(x) = -\nabla_{\theta,\phi}\,\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] + \nabla_{\theta,\phi} D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})]. \tag{5.14}$$

Solving the gradient of the variational loss (second term, 5.14 right-hand side) analytically is straightforward when a) we let the prior to be a centered isotropic multivariate Gaussian $p_\theta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$, and b) we assume a Gaussian form for the true posterior so that we can set the approximate posterior to belong in a Q family of multivariate Gaussian distributions with diagonal covariance matrix, i.e., $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^{\in}\boldsymbol{I})$. In this case, there is no dependency on the gradient parameters θ and ϕ, thus, a closed form of the variational loss exists [120]:

$$\begin{aligned} D_{KL}\big(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z})\big) &= D_{KL}\big(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})) \parallel \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\big) \\ &= -\frac{1}{2}\sum_{j=1}^{J}(1 + \log({\sigma_j}^2) - {\mu_j}^2 - {\sigma_j}^2), \end{aligned} \tag{5.15}$$

where J is the dimensionality of $\boldsymbol{z}$.

However, things are not so simple with the gradient of the reconstruction loss (first term, 5.14 right-hand side), because the negative expected log-likelihood must be estimated with Monte Carlo sampling. Although, a simple Monte Carlo estimator exists with respect to the generative parameters θ, the same is not true with respect to the variational parameters ϕ. The problem with the latter is that we can not move the gradient inside the expectation like this:

$$\nabla_\phi \, \mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] = \mathbb{E}_{z \sim q_\phi(z|x)}[\nabla_\phi \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})], \tag{5.16}$$

because the gradient would be taken with respect to the same parameters as the expectation, i.e., the variational parameters ϕ, and this issue prevents us from obtaining a Monte-Carlo estimation. A score function estimator can solve the problem of placing the gradient inside the expectation in the following way:

$$\nabla_\phi \, \mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] = \mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) \nabla_\phi \log q_\phi(\boldsymbol{z} \mid \boldsymbol{x})], \tag{5.17}$$

but although this estimator can be differentiated, it is known to suffer from high variance, which makes optimization hard to converge.

The reparameterization trick

The AEVB algorithm's key contribution is the introduction of an alternative estimator which is obtained with a very simple but extremely powerful trick, the *reparameterization trick*. The basic idea is to retain the form of the low-variance gradient estimator in Eq. 5.16 by avoiding

the dependence of the expectation on the q distribution and the variational parameters ϕ, so that Monte Carlo estimation is possible. It is accomplished by transforming the random variable $\boldsymbol{z}$, which is drawn from $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$, into a deterministic variable $\boldsymbol{z} = g_\phi(\boldsymbol{\varepsilon}, \boldsymbol{x})$, where $\boldsymbol{\varepsilon}$ is an auxiliary variable, essentially random noise, which is sampled from a simple distribution $p(\boldsymbol{\varepsilon})$ like the standard normal $\mathcal{N}(\mathbf{0}, \mathbf{1})$, and $g(.)$ is a deterministic, invertible and differentiable, vector-valued function with parameters ϕ that is mapping $\boldsymbol{\varepsilon}$ to a more complex distribution. This allows us to rewrite Eq. 5.16 so that the estimator is differentiable since the parameters of the gradient are not the same with the parameters of the expectation:

$$\begin{aligned} \nabla_\phi \, \mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] &= \nabla_\phi \, \mathbb{E}_{\varepsilon \sim p(\varepsilon)}[\log p_\theta(\boldsymbol{x} \mid g_\phi(\boldsymbol{\varepsilon}, \boldsymbol{x}))] \\ &= \mathbb{E}_{\varepsilon \sim p(\varepsilon)}[\nabla_\phi \log p_\theta(\boldsymbol{x} \mid g_\phi(\boldsymbol{\varepsilon}, \boldsymbol{x}))]. \end{aligned} \tag{5.18}$$

The reparameterization trick can be applied only if the latent variables are continuous. This holds because continuous distributions have a sampling property that allows them to be simulated using samples which are drawn from a simpler distribution (independent of the parameters of the initial distribution), and subsequently transformed through a deterministic path $g(.)$ [168]. In the case of a Gaussian approximate posterior (as we have assumed with the VAE), the transformation can be of the form $\boldsymbol{z} = g_\phi(\boldsymbol{\varepsilon}, \boldsymbol{x}) = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and $\odot$ indicates the element-wise product operator. By taking the analytic form of the variational loss in Eq. 5.15, and applying the reparameterization trick on the reconstruction loss, the resulting estimator for a datapoint $\boldsymbol{x}^{(i)}$ is the following:

$$\mathscr{L}_{\theta,\phi}(\boldsymbol{x}^{(i)}) \simeq \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}^{(i,l)}) + \frac{1}{2} \sum_{j=1}^{J} \Big(1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2\Big), \tag{5.19}$$

where J is the dimensionality of $\boldsymbol{z}$, L is the number of Monte Carlo samples, $\boldsymbol{z}^{(i,l)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\varepsilon}^{(l)}$, and $\boldsymbol{\varepsilon}^{(l)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Kingma and Welling [120] refer to this generic estimator as the Stochastic Gradient Variational Bayes (SGVB) estimator, since the resulting gradients can be adapted with stochastic optimization methods such as *stochastic gradient descent* (SGD).

Finally, in terms of implementation, the reconstruction error in Eq. 5.19 is based on the logarithm of the generative model $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ which can be implemented either as a Gaussian (for real-valued output) or as a Bernoulli model (for categorical output). In practice, this means that it can be computed with the Mean Squared Error (MSE) or the

cross-entropy metric, respectively, to measure the difference between the original input and the reconstructed output.

5.3.6 Deep learning perspective

The AEVB algorithm with the SGVB estimator of Eq. 5.19 allows for efficient computation of gradients using backpropagation. Thus, we can use deep neural networks to parameterise and compute the conditional probabilities of the inference and generative models. The inference model $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ is parameterized with an *encoder* network, whose weights and biases are included in the variational parameters ϕ. A second network with weights and biases included in the generative parameters θ, the *decoder* network, is used to parameterize the generative model $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$. In Fig. 5.4, we present a schematic of the components and the computational flow of the variational autoencoder.

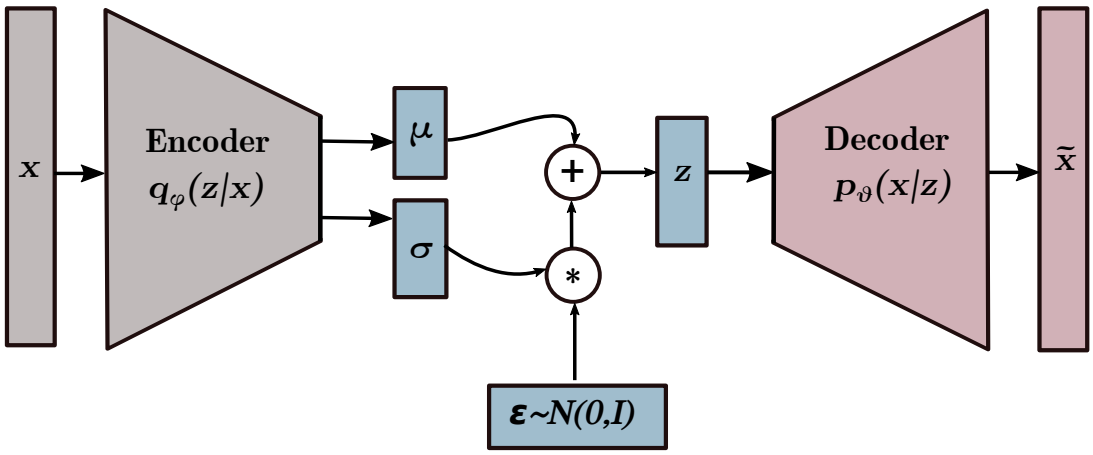

Fig. 5.4 The computational flow of the Variational Autoencoder.

The computational flow in the VAE is the following: during training, at each time step, a sample $\boldsymbol{x}^{(i)} \in \mathbb{R}^K$ of the training dataset is fed to the encoder network which outputs the parameters of a multivariate Gaussian distribution $\boldsymbol{\mu}^{(i)} \in \mathbb{R}^J$ and $\boldsymbol{\sigma}^{(i)} \in \mathbb{R}^J$, with $J \ll K$. Next, the latent encoding $\boldsymbol{z} \in \mathbb{R}^J$ is obtained with the help of the reparameterization trick, by multiplying each element of the $\boldsymbol{\sigma}^{(i)}$ vector with the respective element of a vector $\boldsymbol{\varepsilon} \in \mathbb{R}^J$ drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and adding the product to $\boldsymbol{\mu}^{(i)}$. Then, the latent variable $\boldsymbol{z}$ is fed to the decoder network, which outputs the reconstruction $\tilde{\boldsymbol{x}} \in \mathbb{R}^K$. The reconstruction error and the variational loss are computed, and their gradient is passed with backpropagation to update the weights. After training the VAE, we can remove the encoder network, and only use the

decoder for generating new data. We accomplish this by sampling values $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and feed them directly to the decoder without any reparameterization.

Any neural network architecture can be used to implement the encoder and the decoder of the VAE, and the choice depends on the type of data and the application. The deep learning implementation of the VAE framework can fit a broad range of nonlinear functions and it can scale-up to involve a huge number of parameters and datapoints since we can make parameter updates using small minibatches or even single datapoints. More concretely, for a dataset $\boldsymbol{X}$ with N datapoints, the minibatch version of the SGVB estimator can be constructed as $\mathscr{L}_{\theta,\phi}(\boldsymbol{X}^M) = \frac{N}{M}\sum_{i=1}^{M}\mathscr{L}_{\theta,\phi}(\boldsymbol{x}^{(i)})$, where the minibatch $\boldsymbol{X}^M = \{\boldsymbol{x}^{(i)}\}_{i=1}^{M}$ is a subset of M random datapoints from the complete dataset $\boldsymbol{X}$ of N datapoints.

The stochastic power of the VAE

The structure of the inference and generative model implemented with neural networks is, in essence, an autoencoder, but as we have already mentioned the big difference between a VAE and a standard autoencoder is in the optimisation objective. Standard autoencoders optimisation is based on the reconstruction error, but the fact that they lack a regulariser prevents them from learning meaningful representations of the data [23]. In contrast, this regulariser is the variational loss in the SGVB estimator used by the VAE. Furthermore, the VAE bottleneck defines a stochastic latent space which can be used, after training, to draw samples which can be used to generate new data. In the standard autoencoder, the bottleneck is just a compression layer with deterministic output, and it cannot be sampled stochastically.

5.3.7 Posterior collapse

A well-documented problem arising in the training of VAEs is the *posterior collapse*. That is when the model instead of learning meaningful latent features from the input, learns to ignore the latent variables resulting in a posterior distribution that collapses to the prior (usually selected to be an uninformative prior, e.g., a unit Gaussian). More concretely, posterior collapse indicates that the training is trapped in a trivial local optimum where $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \simeq p(\boldsymbol{z})$ for all $\boldsymbol{x}$. Empirically, the problem can be detected during training when the variational loss $D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})$ becomes almost zero.

A lot of recent efforts to understand and alleviate the problem have been reported. Many of these efforts focus on modifications of the ELBO objective by controlling the impact of the KL divergence term $D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z})]$. Bowman et al. [33] proposed the *KL cost annealing* approach, in which they multiply the variational loss with a variable weight that increases progressively from 0 to 1 during training. Intuitively, they begin with a

standard autoencoder, since only the reconstruction error matters at the beginning, and they progressively change to a VAE. This approach is meant to force the model to learn encodings which enable a faithful reconstruction and then gradually smooth out the latent space by using the prior constraint. Unfortunately, it appears that in practice the KL annealing is not sufficient after the weight exceeds a certain value, and as it approaches a value equal to 1, the posterior collapses again.

Higgins et al. [96, 43] propose the β-VAE modification of the objective in which they rescale the KL divergence term with a fixed β coefficient which is treated as a hyperparameter. The coefficient is derived as a Lagrangian multiplier and the larger it is, the stronger the pressure on the approximate posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ to match the prior $p_\theta(\boldsymbol{z})$. This work focuses mainly on learning disentangled latent representations, i.e., underlying factors that affect the disjoint properties of the data representation [23]. The authors postulate that this can be encouraged with $\beta > 1$, but in practice, although the β coefficient allows for better control of how much of a constraint we want to impose on the latent distribution, choices of $\beta \gg 1$ often come with a cost on the reconstruction fidelity since there is a tradeoff between reconstruction and variational loss.

Alemi et al. [1], examined this tradeoff between the reconstruction accuracy and the latent space compression from an information theory point of view, by deriving rate-distortion curves based on the mutual information between the observed data $\boldsymbol{x}$ and the latent values $\boldsymbol{z}$. Their experiments demonstrate the tradeoff and show that the posterior collapse occurs in architectures that use a powerful decoder. They propose a straightforward solution, namely to reduce the KL divergence β coefficient to $\beta < 1$.

Powerful decoders—as in the case where an expressive autoregressive network is used to implement the decoder—have received much attention as a potential cause of the posterior collapse problem. The phenomenon appears to be prevalent in these implementations, and often weakening the decoder's capacity has been proposed as a solution [48]. Nevertheless, it has been shown recently by Lucas et al. [146] that posterior collapse may occur even without powerful decoders when testing simple linear VAEs.

Posterior collapse causes and solutions attract much interest since this optimisation challenge is a limitation for using more expressive VAEs, especially in representation learning.

Chapter 6

Generating robotic EBL with a Variational Autoencoder

6.1 Introduction

This chapter describes the design and implementation of a Variational Autoencoder (VAE) for the generation of robotic Emotional Body Language (EBL) animations. The model is trained with a small dataset of EBL animations and learns a latent representation of postures which can be sampled and interpolated to generate new previously unseen EBL animations. Furthermore, we conduct an exploratory analysis, the aim of which is to gain insight into the properties of the latent space and the model's capacity to generate many different variations of robotic EBL. The study described in this chapter was published in [153].

The chapter is organised as follows. The first section describes the animation set and the preprocessing of the data to create a training set, the VAE model implementation in terms of architecture and training parameters, the sampling methodologies and the software we used. In the second section, we present and discuss the results of the exploratory analysis, and in the third and final section, we draw our conclusions and discuss the limitations of this study and the next steps.

6.2 Methods and materials

This section presents the dataset, the VAE implementation, and the methodologies we used to sample the model and generate animations.

6.2.1 Dataset

We begin with a description of the animation set and how it was preprocessed to be used in the VAE training.

EBL animation set

In the following experiment, we trained a VAE model with 36 robotic animations selected and evaluated during our first experiment described in Chapter 4. Initially, the animations incorporated information from three modalities: motion, eye LEDs colours, and non-linguistic sounds. For this experiment, we removed all the information except motion. Consequently, each animation is a non-fixed length sequence of keyframes, each describing a posture configuration along with the timestamp indicating when the posture must be reached. During execution, the keyframes are interpolated with cubic Bezier interpolation applied locally at each joint, to get the angles for the intermediary frames, a method called *inbetweening* in graphical animation. More concretely, a keyframe is represented with a vector of 17 real values describing the angles of Pepper's 17 joints (Fig. 6.1) in radians and one value for the timestamp in seconds indicating when the particular configuration of angles should be reached. We did not include any information about the emotional content described either by categorical or dimensional tags (valence and arousal) since the experiment will follow an unsupervised learning approach based exclusively on the motor content.

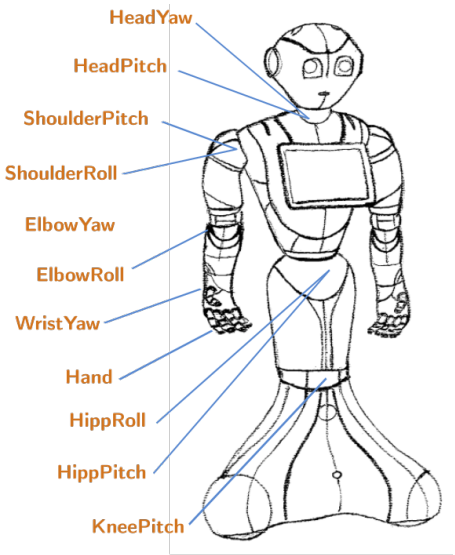

Fig. 6.1 The angles of the 17 joints of the Pepper robot define a posture. The current configuration is the StandInit posture.

Preprocessing

The animations' keyframe representation described above is the original one used for the design of the animations and their execution through the robot's NAOqi operating system. However, for the VAE training, we had to obtain a data representation that is more efficient for deep learning. We needed to maximise the training set samples' size and keep their content as succinct as possible, without unnecessary information. We figured out that the best way to achieve this is to use a constant sampling rate for capturing the intermediary postures. This approach also allows us to remove the dimension of the timestamp since it would be redundant.

To obtain this transformation of the dataset, we had the animations executed by the real robot, and we used the NAOqi built-in function *ALMotionProxy::getAngles* to record the positions of the angles with a sampling rate of 20 frames per second. Before the recording, we made a small modification in the keyframe representation by adding an extra keyframe describing the StandInit posture (see posture in Fig. 6.1) at the beginning and the end of each animation (whenever it was missing), to obtain a clear onset and offset of an animation. In total, the initial keyframe representation dataset contained 697 keyframes, and with up-sampling, we obtained 4696 posture frames, i.e., vectors of joints' angles.

The next issue that concerned us in preparing the training set was related to the prevalence of possible body orientation bias. For example, if many postures assume a right-side orientation, the training could be heavily biased, and the network might not learn enough postures with left side orientation. Asymmetries in facial expression [31] or head motion [75] have been associated with the lateralisation of emotion processing in the brain, and similarly this could be the case with body posture. There are two main hypotheses proposed: the Right-Hemisphere Hypothesis, which supports the dominance of the right half of the brain is uniquely specialised in emotion, either positive or negative [30], and the Valence-Specific Hypothesis, which posits that the right hemisphere is specialised in negative emotions and the left hemisphere in positive emotions [59]. However, there are also studies supporting that the two hypotheses are not always mutually exclusive [185, 118]. We decided to assume that the interpretation of the bodily expression of emotion is independent of left or right side orientation with respect to the body's vertical axis. This assumption allowed us to augment the dataset by having every posture mirrored from left to right or the opposite. Nevertheless, we avoided any mirroring with respect to the horizontal axis, e.g., changing upward motion to downward or vice versa, since such movements are often found to correlate with emotional expression in studies of bodily affect [56, 53]. More specifically, to create a mirrored posture we swapped the values of the entire chain of arm joints between the left and right arm, i.e., shoulder pitch and roll, elbow roll and yaw, wrist yaw, hand. For the arm's roll and yaw

joints, we had to invert the signs of their values because the ranges of these joints had inverse signs between left and right. Furthermore, we also mirrored the head yaw and hip roll joints since they are also moving rightward and leftward with respect to the body's vertical axis. For this mirroring, only the signs needed to be inverted. After the mirroring augmentation, the size of the training set increased from 4696 to 9392 posture frames.

The last step in the preprocessing pipeline was the data normalisation. The ranges of plausible joint values differ among the joints; thus, we rescaled all the values in [0, 1] with Min-Max Normalization. We shuffled the frame samples and used 80% for the training, and the rest for the validation.

6.2.2 The VAE network implementation

This section describes the VAE model implementation in terms of network architecture, training parameters and configuration.

Architecture

For the implementation of the encoder and the decoder of the robotic EBL generation VAE, we used two Multilayer Perceptrons (MLP) of equal capacity. Each has three dense layers with 128, 512, and 128 units, but the decoder MLP has an additional dense layer after the output, serving as a reconstruction layer, with 17 units to match the input dimensions (Fig. 6.2).

An MLP is a feedforward neural network with several fully connected layers. The MLP was a simple and adequate choice for the problem since the input has only 17 dimensions. A Convolutional Neural Network (CNN), which is more efficient for high dimensional pixel spaces in image classification, would be an unnecessarily more complicated choice. Also, it could be argued that a better alternative might have been to use a recurrent neural network (RNN) since it can capture sequence dynamics. However, recall that we train our VAE with shuffled postures instead of sequences and that the VAE's objective aims to learn a continuous latent manifold that can be interpolated and generate trajectories that appear sequential and smooth; thus, we can achieve the sequential output even with the MLPs.

Although the choice of dense layers with 128 to 512 units might not appear particularly lightweight, it was decided after several experiments in which we tested implementations with less or lighter layers and unequal capacities between the encoder and the decoder. The selected setup was reasonably fast to train on a CPU with 6 multiprocessors (approximately 20 minutes to complete training), stable and it converged to low reconstruction and validation error.

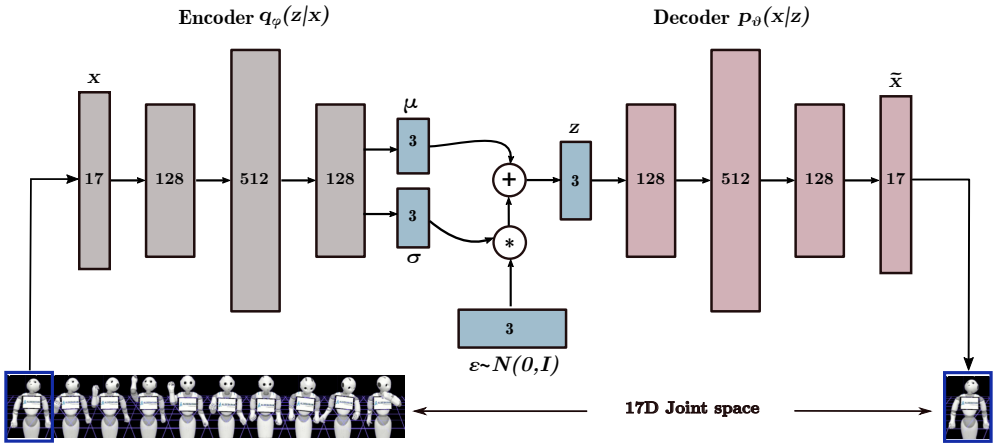

Fig. 6.2 The architecture of the robotic EBL generation VAE. The orthogonal blocks represent network layers with their numbers indicating the units of each layer. At each step, a posture is fed as an input vector $\boldsymbol{x}$ of 17 values, and it is reconstructed in the output $\tilde{\boldsymbol{x}}$.

For the latent layer, we decided to use a strong bottleneck of 3 dimensions. We also tested higher dimensions, but we did not notice any improvement in the reconstruction and validation error. The latent layer infers the parameters of a multivariate Gaussian distribution (a 3D Gaussian in our case), so it consists of two sub-layers, one for the latent mean and one for the latent standard deviation (Gaussian parameters μ and σ), with 3 units each (Fig. 6.2). The output of these layer is used along with a 3D sample $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ to produce a latent 3D vector $\boldsymbol{z}$ which is then fed to the decoder (reparameterisation trick).

We used a layer of rectified linear activation functions (ReLU) after each hidden layer, while linear activations followed the output layers. The decoder's reconstruction layer uses a sigmoid function to rescale the output in the range [0,1]. After each hidden layer, we use a dropout layer, which randomly sets half of the units to zero [98]. At first glance, this choice might appear redundant since the VAE already uses a regulariser in its loss function (the variational loss), but in practice, the dropout appears useful because it decreases the computational cost without making training less effective. Finally, we trained for 200 epochs. The architecture specifications are summarized in Table 6.1.

Training parameters and configuration

The training parameters are summarised in Table 6.2. The network weights were initialised with a Xavier uniform initialiser [82], i.e., each unit's initial weight was sampled from a uniform distribution with limits normalised by the sum of its input and output units. For the

Table 6.1 Robotic EBL VAE architecture specifications

	Encoder	**Decoder**
Input	$\mathbb{R}^{17}$	$\mathbb{R}^{3}$
Reconstruction	NA	$\mathbb{R}^{17}$ (sigmoid)
Hidden layers	3 FC layers	
Hidden units	128, 512, 128	
Hidden activation	ReLU	
Output activation	Linear	

Table 6.2 Robotic EBL VAE training parameters

Training parameters
Xavier initialization
Dropout
Adam, $lr = 1e-4$
$\beta = 0.001$
Batch size: 32
Epochs: 200

optimisation, we used an Adam optimiser [119] with a learning rate equal to 1e-4. We used the Mean Squared Error (MSE) between the original frame and the reconstructed output to compute the reconstruction error. The variational loss was computed in the analytic form (Eq. 5.15). We sum the two losses for each input sample and we average across minibatches of size equal to 32.

The most challenging hyperparameter to tune was the β coefficient of the variational loss. With a $\beta = 0$ the model is reduced to a standard autoencoder, which can compress the input through the 3D latent space, and reconstruct it faithfully, but since it is a non-probabilistic network it can not be a generative model. In the original VAE framework, there is no scaling of the variational loss ($\beta = 1$), while in the β-VAE network [96] it is proposed to use $\beta > 1$ for better disentanglement of the latent factors underlying the data distribution. We tested both alternatives for β, but in both cases, we obtained poor results because of posterior collapse. The latent space was completely shrunken to the unit Gaussian prior and was ignored by the decoder who was unable to reconstruct the input. We found that the problem could be alleviated by decreasing β below 1, and we empirically ended it up assigning $\beta = 0.001$.

6.2.3 Sampling the latent space

The sampling process is applied directly to the 3D latent space by drawing samples from a 3D Gaussian. Subsequently, the samples can be interpolated to produce a latent 3D trajectory that defines a latent animation. At this point, it is important to note that the latent space, besides encoding the configuration of postures, also captures the temporal dynamics that define the transitions between postures. We will now briefly overview different interpolation methods that can be used. We will also describe the method we followed to sample latent trajectories in a systematic way to examine what their properties are and how they decode into the robot's joint space.

Interpolation methods

Since we want to be able to generate animations in the 17D joint space, we need to sample latent trajectories instead of just individual latent datapoints. This can be accomplished by applying interpolation between the sampled latent datapoints. Interpolation or inbetweening methods are well studied and widely applied in computer animation for filling in the intermediary frames between keyframes [91]. As we have already mentioned in the description of our original animation set, the training set animations were created with pose-to-pose design of the keyframes and then cubic Bezier interpolation was applied locally at each joint, to derive the intermediary postures. The main difference with the latent space interpolation is that it is applied globally instead of individually at each joint, since all the joints information are compressed into three latent dimensions.

Different interpolation techniques can be used, and here we examine three of them to find out how they affect the decoding of the latent trajectories. The first one is the *linear interpolation* (Lerp), which is very simple to understand and apply, and is the most common choice for sampling generative models. The Lerp interpolation produces a straight line connecting the endpoints. The distance between successive intermediary points on this straight line is equal. For more than two datapoints, the pieces of the linear interpolates are connected. However, although it results in a continuous trajectory, this trajectory is not smooth.

The second interpolation technique we examined is the *spherical linear interpolation* (Slerp) [205]. In Slerp, the segments connecting the endpoints are curved lines. The intermediary points are non-equidistant, with more points accumulating near the endpoints, and less toward the segment's middle. This creates a slow-in and slow-out effect from one endpoint to another. As is the case with the Lerp interpolation, the overall trajectory is not smooth, since the derivative is discontinuous at the connection points between segments.

White et. al [226] suggest Slerp as a more appropriate method for sampling latent interpolants, because it accounts for the curving shape of the latent space, and is less prone to traverse blind spots, that is, latent space locations far away from the prior.

Finally, we examined B-splines interpolation [15], which allows to smoothly join together adjacent segments to produce an overall smooth curve. More specifically, we used a cubic spline interpolation with a *natural boundary* condition, which applies low-degree polynomials piece-wise for every segment, and enforces zero second derivatives at the endpoints, resulting in less oscillations compared to polynomial interpolation, with enough stiffness around the endpoints that ensure a more natural transition.

In summary, we will examine how the decoded latent trajectories are affected by the interpolation method used. We will apply three different interpolation algorithms: 1) Linear interpolation (Lerp), 2) spherical linear interpolation (Slerp), and 3) B-spline interpolation. Lerp produces continuous but not smooth trajectories, comprised of straight lines with equidistant points. Slerp produces continuous but not smooth trajectories, comprised of curved lines which accumulate more points at their ends. B-splines produce continuous and smooth trajectories.

Spherical grids

The methods we discussed so far for sampling latent trajectories, pertain to random sampling of latent datapoints and then interpolate them to create a latent trajectory. It would be more informative to use a more systematic way to sample latent trajectories and examine how they decode into the 17D joint space. We are particularly interested in exploring the granularity potency of the latent space, and examining if there are properties that we can use to generate more targeted animations in terms of their emotional interpretability.

Since we used a 3D Gaussian prior to constrain the latent variables during training, the latent space distribution can be visualised as a spherical ball. Thus we can use topological features of the sphere, such as the radius, to extract trajectories systematically and explore the properties of the latent representation and their impact in the decoding process. Based on this motivation we conceptualised the idea of projecting spherical grids on the latent space and then interpolate along their longitudes to derive latent interpolants (Fig. 6.3).

We defined spherical grids of different radiuses ranging from 0.5 to 10, for 10 different values. For each grid we defined 10 longitudes which are spread equidistantly on the sphere. Each longitude contained 20 points, which were interpolated with Slerp of 10 steps at each segment to produce latent trajectories of 200 frames. Furthermore, each grid was rotated in three different ways, so that the axis was parallel to one of the three latent dimensions LD1, LD2, and LD3. In total, for the 3 latent dimensions, the 10 radius values, and 10 longitudes,

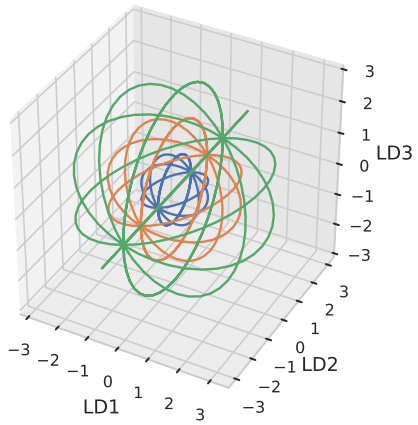

Fig. 6.3 Three spherical grids with their axes parallel to the latent dimension LD2 and radiuses ranging from 1 (blue) to 3 (green).

we got 300 latent trajectories of 200 frames each. As an example, in Fig. 6.3 we plotted 3 such grids, with radiuses equal to 1, 2, and 3, with their axes parallel to LD2. Each of these grids has 10 longitudes and every longitude defines a latent trajectory of 200 latent postures which can be decoded into an animation.

6.2.4 Generating animations

For the generation of new animations we fed the 3D latent trajectories to the decoder network, sample by sample, and obtained a 17D trajectory of postures, to execute as an animation with the NAOqi built-in functions.

6.2.5 Software

For recording and executing robotic animations with the real Pepper robot we used NAOqi 2.5 SDK[1]. For the robot simulations we used the Choregraphe Suite[2]. The code for the VAE model was adapted from the TFModelLib collection of neural networks[3]. For the implementation of B-splines we used Splipy[4]. The code for Slerp interpolation was taken from the Plat repository[5]. The rest of the code for the data preprocessing, sampling, spherical grids, visualizations, etc. was written in Python 3 with packages such as NumPy [178], SciPy [219], pandas [162], scikit-learn [182], Matplotlib [104].

[1]NAOqi 2.5: http://doc.aldebaran.com/2-5/home_pepper.html
[2]Choregraphe Suite 2.5: http://doc.aldebaran.com/2-5/software/choregraphe/index.html
[3]TFModelLib repository: https://github.com/nhemion/tfmodellib
[4]Splipy v1.3.1: https://sintefmath.github.io/Splipy/
[5]Plat repository https://github.com/dribnet/plat

6.3 Results

In this section we present the training results and the exploratory analysis we conducted regarding the sampling and interpolation of the latent space. We investigate the trajectories of the generate animations in terms of their variability. We derive a hypothesis on how we can use the 3D latent space's geometrical features to condition the generation of animations with the arousal component.

6.3.1 Training performance

The VAE was trained for 200 epochs, and it converged to a training error equal to 0.007, and validation error 0.0071 (see Fig. 6.4). At test, the reconstruction error was 0.0028, and the variational loss was 4.2175 (without the β scaling). In Fig.6.5 we visualise all the postures of the dataset encoded in the 3D latent space, colour-coded according to the animation they represent. Each datapoint is an encoded posture. It can be seen that the VAE learned a tight latent space that spreads similarly in every dimension, resembling a 3D Gaussian with a mean of zero. Now, sampling from a 3D Gaussian—even from areas that have not been encoded during training—, interpolating the samples and decoding the latent trajectories will generate new, realistic postures with high probability .

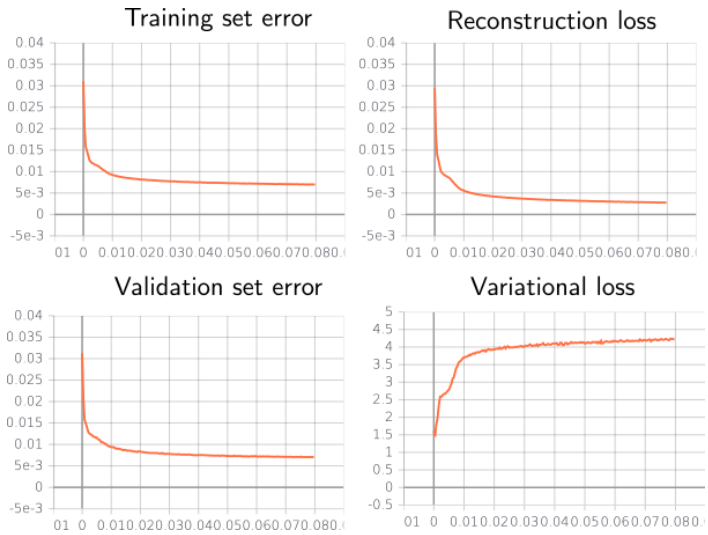

Fig. 6.4 The training and validation set errors. Reconstruction and variational losses are presented for the validation set.

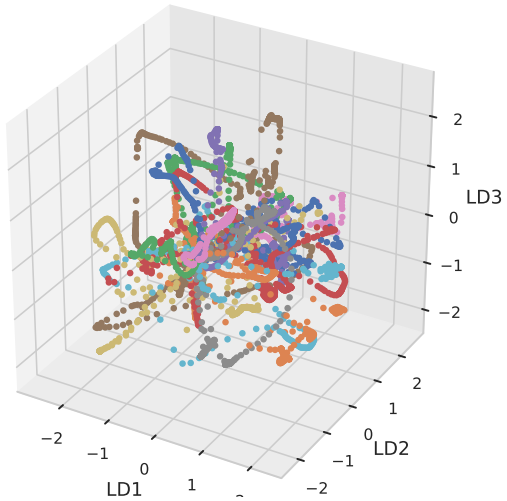

Fig. 6.5 The training set animations encoded in the latent space after training the VAE network for 200 epochs. The colour code represents different animations. The latent space is defined by three latent dimensions: LD1, LD2, and LD3.

6.3.2 The effects of different interpolation methods

To explore how the three different latent space interpolation methods affect the generated animations we created a small animation of six frames in the 17D joint space, by randomly selecting four posture frames from the dataset, and adding a frame of the StandInit posture at the beginning and the end. Then, we encoded the six frames in the latent space and we interpolated them with 100 evaluation points per segment for Lerp and Slerp interpolation, and 500 points for B-spline globally, deriving three latent interpolants of 500 datapoints each. The latent interpolants are plotted in Fig. 6.6. They all begin and end with the latent encoding of the StandInit posture, and they all pass through the four intermediary postures. For Lerp and B-spline the trajectories are curved, but B-spline is the only one that is completely smooth even at the connecting points.

To examine the interpolation method's effects in the joint space, we decoded the three latent interpolants with the VAE decoder into the 17D joint space. To visualise and compare the three animations, we averaged each of them across the 17 joints. The averaged trajectories are presented in Fig. 6.7, where we compare them against a fourth trajectory (black dashed line). This trajectory represents an animation with the same six keyframes interpolated directly in the 17D joints space with the NAOqi Bezier interpolation method, similarly to the process used by the animators who created the original animations.

It can be seen in Fig. 6.7 that the decoded VAE trajectories share a common trend with the NAOqi trajectory, but they exhibit far more complexity. The NAOqi interpolation is applied locally at each joint, without taking into account the trajectories of the other joints.

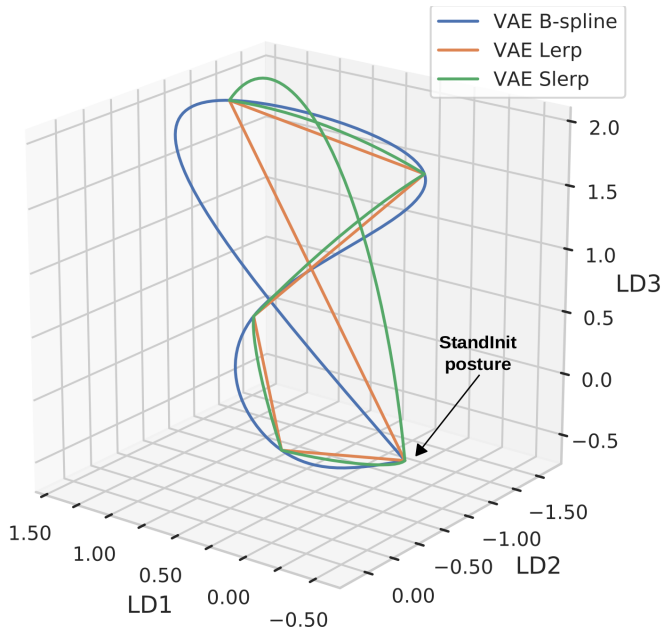

Fig. 6.6 Different latent interpolation methods for the same sequence of postures. StandInit is the initial standing posture of the robot. All interpolants are passing through the key postures, but they follow different trajectories in doing so.

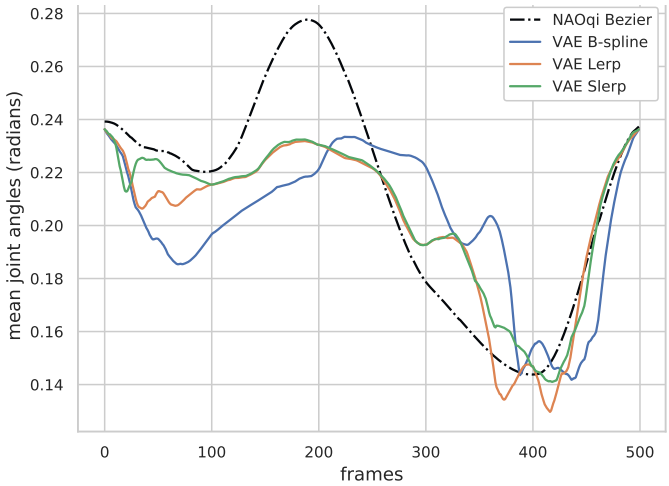

Fig. 6.7 Three animations decoded from the latent interpolants of Fig. 6.6 and averaged across joints, compared to the Naoqi interpolation applied directly in the joint space (black).

In contrast, the VAE generated animations originate from a global interpolation over the latent encodings. Thus, it encompasses information from all the joints and their interactions. Furthermore, the variation apparent in the VAE generated trajectories can be explained by the fact that the latent space interpolation traverses areas which encode many different postures, not just the keyframes used to construct this 6 frame sequence, causing many fine-grained variations to be injected in the generated features.

Finally, by comparing the VAE generated trajectories to the different latent space interpolation methods, we observe that Lerp and Slerp overlap a lot, while the B-spline exhibits a delay, possibly because it is not applied piece-wise. Although it passes through the six keyframes, it does not use the same number of evaluation points per segment, so keyframes appear shifted. In the following analysis we will be using Slerp interpolation.

6.3.3 The animations generated from the spherical grids trajectories

Instead of random sampling, we used the spherical grids templates described in Section 6.2.3 to systematically generate 300 animations. We present these animations in Fig. 6.8 as single trajectories, after averaging each over the 17 joints, so each trajectory in a block is an animation. Each of the three horizontal panels contains the trajectories of the animations generated from spherical grids with their axes parallel to one latent dimension (LD1 to LD3). The ten blocks in each panel represent the 10 longitudes of a spherical grid. The colour of the trajectories in a block indicates the different radiuses used to scale the spherical grids. For example, the black trajectory in the left topmost block represents an animation decoded from a spherical grid's first longitude with a radius equal to 0.5 and axis parallel to LD1. The visualisation of the 300 animations demonstrates the latent space's potential concerning the granularity of the encodings. Furthermore, the visualisation highlights the topological properties that can be exploited to generate animations of a targeted emotional content.

Granularity in the generated animations

Fig. 6.8 shows that every trajectory is different from the rest. This exhibits the latent space's potential to generate numerous different animations, even when it is trained with just 36 animations. We can observe the variation between the blocks corresponding to different longitude lines, the variation with respect to the radius scaling, and the variation with respect to the spherical grids' axis orientation. The variation occurs within each latent dimension, as well as between different latent dimensions. This sample of 300 generated animations demonstrates the granularity that is inherent in the latent space. Even by just reversing the frames' sequence, we can double the variations while keeping the smoothness and sequential

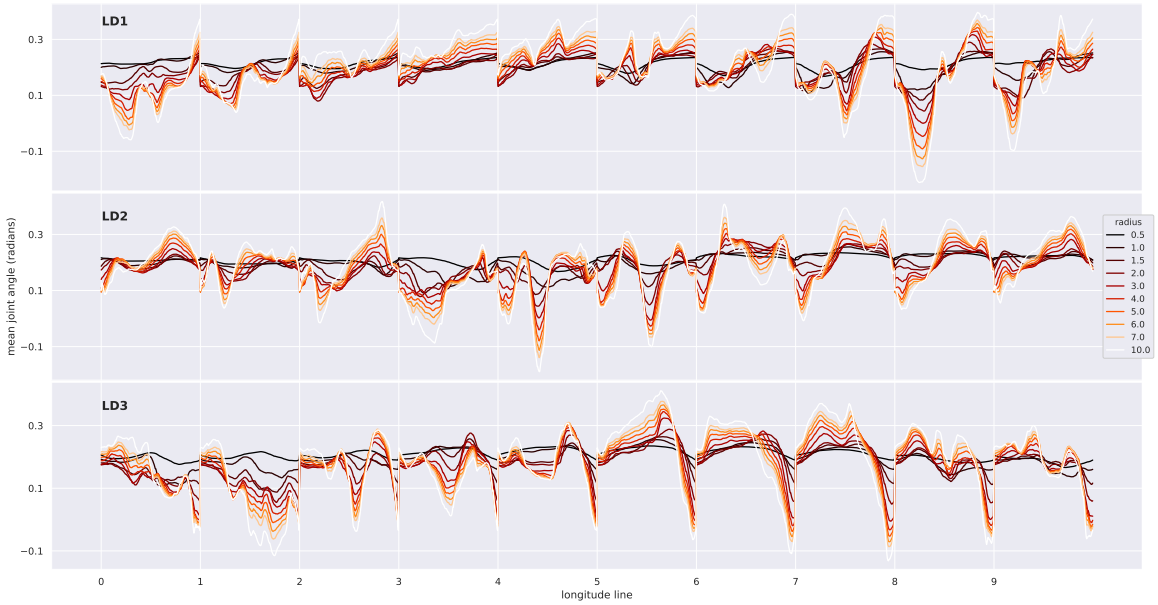

Fig. 6.8 Decoded interpolations in the joint space, averaged across joints. Each of the three panels represents the animations that were decoded from the longitudes of spherical grids projected on the latent space with their axes along one latent dimension (LD1-LD3). For each latent dimension, the spherical grids were scaled with radiuses ranging from 0.5 to 10, producing ten animations for each longitude. The plots demonstrate the granularity of the animations sampled from the latent space of the VAE model and the fact that the motion is amplified as we move further away from the core of the latent space. We suggest that the later feature can be used to model the arousal dimension of emotion.

characteristics. It becomes apparent that the robotic EBL VAE can generate numerous animations of different content very easily.

The amplitude of motion is modulated by the radius

Another observation that becomes apparent by inspecting Fig. 6.8, is the increase of the trajectories' amplitude with respect to the radius scaling. This effect is related to the latent topology and it indicates that we can sample subtle emotional expressions near the core of the latent space, while as we move outward by increasing the radius, we can sample latent trajectories that will decode into expressions of heightened activation. Therefore, we could potentially use this progressive alteration of the amplitude encoded in the latent space topology to manipulate the arousal dimension of emotion in the generated animations.

It should be noted that this result is not possible—or at least it would not produce a smooth effect—if we would just manually scale the joints values. Such manipulation applied directly and globally at the joints space would require extremely cumbersome tuning, otherwise,

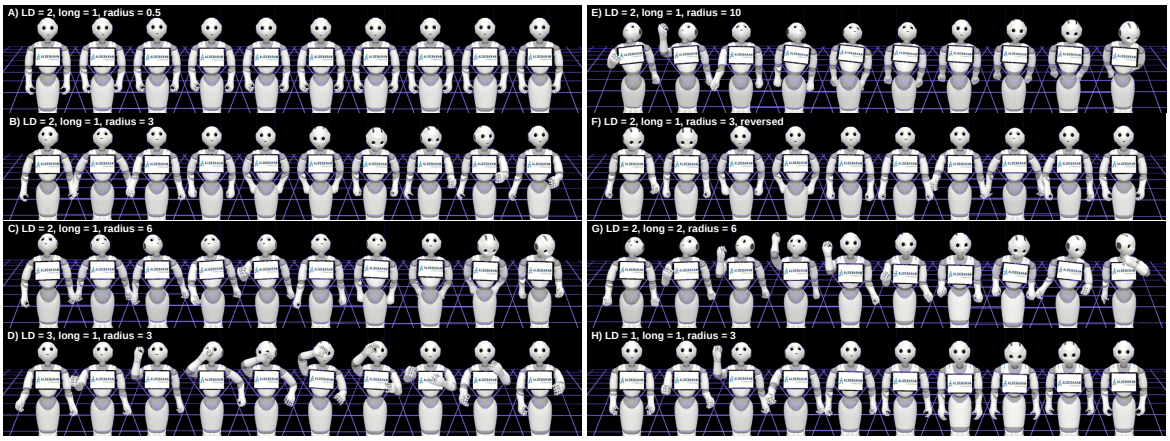

Fig. 6.9 Simulation of selected generated animations. The abbreviations are LD = latent dimension and long = longitude line. A: Very subtle motion is sampled from the core of the latent space. B-C: Moving outward from the core, the motion has higher activation. D: A different animation is obtained when rotating the spherical grid axis to another latent dimension (compared to B). E: Sampling out of the limits of the learned latent space causes unnatural motion artifacts. F: Reversing animation B gives a different expression. G: Adjacent longitudes decode into a different animation (in comparison with C). H: Changing the grid axis to another latent dimension generates different animation (in comparison with B and D).

it would induce artifacts in the smoothness of the motion and result in unnatural postures, because of the dependencies between different joints and the different ranges of values they can take. Consequently, we believe that with our trained VAE model we have obtained a feature for smooth manipulation of the motion amplitude. This feature is remarkably easy to control and appears promising for modifying the arousal content of an animation. It can be configured by changing the radius of the spherical grid template, or the standard deviation of a 3D Gaussian distribution in random sampling.

6.3.4 Display on the robot simulator

In Fig. 6.9 we demonstrate the outlined effects by displaying the generated animations on the Choregraphe Suite robot simulator for Pepper. By comparing sequences A, B, C and E, it becomes apparent how the radius of sampling can modulate the expression's amplitude. These four sequences were decoded from the same latent longitude of spherical grids with axes parallel to LD2. The only difference between them is the radius of the grids. Starting with a radius equal to 0.5 in A, the motion is very subtle and the robot appears almost as not moving. Then as we increase the radius to 3 for sequence B, and 6 for sequence C the motion becomes progressively more activated.

For sequence E, the radius is equal to 10, and it can be noticed that some unnatural joints configurations become apparent, e.g., the elbows are bent backwards. This is because sampling at this radius exceeds the limits of the learned latent space. Recall that we trained the VAE with a variational loss that forces the latent distribution to resemble a unit 3D Gaussian prior, spreading spherically within a unit around a zero mean. However, this loss did not converge to zero, thus the learned distribution is not identical to the prior and has a higher spread. As shown in the latent space visualisation 6.5, this spread is at least equal to 3 for the encoded dataset. So we have a pretty good idea about the limits of the latent space, and by testing a bit higher radiuses than 3, we easily detect them and avoid generating animations with such artifacts. In our VAE model, we detected this limit effortlessly after a couple of tests at 7 units, so by restricting the radius or the standard deviation below 7, we can be sure that we will be generating realistic animations.

The variation with respect to the latent longitudes is demonstrated by comparing sequences G and C; adjacent longitude lines from the same grid, i.e., same radius and orientation, decode into different animations. Similarly, the comparison of sequences D and H pinpoints the variation due to the rotation of the spherical grids' axes when the radius and longitude line remain the same. Finally, sequence F in comparison with sequence B illustrates another variation that can be obtained by just reversing the sequence's frames.

6.3.5 Display on the real robot

Finally, we tested the generated animations on a real Pepper robot to examine how they are executed in real-world conditions. A video with the physical robot executing several generated animations with the properties outlined above is available online[6]. It can be noticed that many of the animations presented in the video, which are generated from spherical grids with the same orientation with respect to the latent dimension, they appear to finish abruptly and in the same or very similar posture. These effects are because latent longitudes end up at the same latent encoding located at the spherical grid's pole. This point encodes a posture at which we arrive smoothly, but since we do not continue to interpolate, the motion appears to be halted abruptly. We discuss this further in the following section.

6.4 Discussion and conclusion

This chapter presented the implementation and application of a VAE model for the automatic generation of synthetic robotic EBL animations. The network is trained with a small set

[6]VAE generated animation set video: https://youtu.be/sdsp05rqcJA

of high-quality robotic animations of emotional content and can generate numerous new variations that appear realistic and expressive. We propose this approach as an alternative to hand-coded, pose-to-pose animation techniques which are cumbersome to design and thus limited in number and variations. The VAE generated animations can be of great value in social and affective robotics applications where we need a lot of different animations with fine-grained variations to sustain a robot's 'illusion of life' in long-term human-robot interaction. Such an effect cannot be accomplished with small animation sets since the expressions will become repetitive and predictable after a point.

Furthermore, we conducted an exploratory analysis to better understand the properties of the latent space of the model and their impact on the decoded animations. Interestingly, we discovered that the motion amplitude in the generated animations can be modulated by the radius or the standard deviation of the sampling applied to the spherical latent space. Assuming that the amplitude of motion can be an aspect of heightened activation, we suggest that this feature can be potentially used to model the arousal dimension of the emotional content in the generated animations. This hypothesis remains to be tested further in a user study.

6.4.1 Limitations and next steps

This first attempt to generate animations for Pepper with a VAE has been quite conducive to better understand robotic EBL synthesis, but it also revealed several shortcomings that we need to address in the next steps. The most important limitation lies in the fact that the proposed model can not generate animations of prespecified emotional content. Although, we ensure that the generated animations will encompass some emotional trait—given that we trained the VAE with examples of EBL—, we can not control the generative process to output emotions of specific valence and arousal. A socially responsive robot needs not only to express emotion but also to express the appropriate emotion with respect to the situation. Therefore we need to be able to condition the generative process to specific valence and arousal values. As reported previously, we have a potential candidate feature to condition the generated animation in terms of arousal, but a solution is also necessary for valence conditioning. To address this crucial requirement, in the next chapter we will introduce the Conditional Variational Autoencoder (CVAE), a slightly modified version of the VAE that allows us to condition the generated output with some given label.

The second important limitation of the present work is the lack of human evaluation of the generated animations. We trained a model with EBL animations, and technically, the model converged and learned these features. However, to ensure that the generated animations are interpreted as involving emotional traits, a user study is indispensable. We avoid conducting

a user study at this stage, because we wanted to complete the conditioning requirement first and then collect evaluations on the interpretability of the valence and arousal conditioning.

A technical shortcoming that we want to address in our next effort is the abrupt ending at the same posture when using the spherical grids longitudes to interpolate the latent space. We would like to be able to systematically sample latent trajectories, but in such a way so that they have a smooth and clear onset and offset.

Finally, the original animations contained more information than just the motion sequences, e.g., eye LEDs patterns. In the next steps, we attempt to take advantage of this information to enhance the robot's expressivity.

Chapter 7

Generating robotic EBL of targeted valence and arousal

7.1 Introduction

In this chapter, we outline our efforts to address the previously mentioned limitations(Section 6.4.1) and obtain a complete pipleline that generates targeted emotion animations by conditioning the process with specific valence and arousal values. The proposed pipeline also aims to improve the generated animations' expressiveness by adding a second modality besides motion, and in particular, the eye LEDs colour sequences. Part of the work presented here is published in [151]. The full dataset we used for training the CVAE is available[1].

The chapter is organized as follows: The first section describes the theoretical aspects of the Conditional Variational Autoencoder which is the model we will apply. Afterwards, we outline the methodologies we used for restructuring the training set, implementing and sampling the model.

7.2 The Conditional Variational Autoencoder

The VAE is a very powerful generative model but in its standard form it does not allow any control on the generation process in terms of enforcing specific attributes on the output. This means that after we have trained the model, the output can be of many different classes. Controlling the generative process so that the decoded output matches some given attribute can be very important in certain applications. The Conditional Variational Autoencoder (CVAE) [208], an extension of the VAE framework [120, 110, 121], is trained to maximize a lower

[1]REBL-Pepper Dataset: https://github.com/minamar/rebl-pepper-data

bound on a *conditional* data log-likelihood, and therefore it learns to make a discriminative prediction additionally to reconstructing the input. More concretely for a datapoint $\boldsymbol{x}$ and an auxiliary variable $\boldsymbol{c}$ that holds a ground-truth label or an attribute, the ELBO from Eq. 5.10 is modified as follows:

$$\log p_\theta(\boldsymbol{x} \mid \boldsymbol{c}) - D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{c}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{c})] = \mathbb{E}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{c})] - D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{c}) \parallel p_\theta(\boldsymbol{z})] \tag{7.1}$$

where the expectation is taken over all $z \sim q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{c})$. The new objective is thus to maximize the data likelihood conditioned on $\boldsymbol{c}$. It should be noted that $\boldsymbol{c}$ can be discrete or continuous. The loss function to be minimized is the following:

$$\mathscr{L}_{\theta,\phi}(\boldsymbol{x}, \boldsymbol{c}) = \underbrace{-\mathbb{E}_{z \sim q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{c})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{c})]}_{\textit{reconstruction error}} + \underbrace{D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{c}) \parallel p_\theta(\boldsymbol{z})]}_{\textit{regularization term}}. \tag{7.2}$$

The modified CVAE loss function conditions the generative process on a given label or attribute $\boldsymbol{c}$. During training, at each step, we concatenate input $\boldsymbol{x}$ and label $\boldsymbol{c}$, and we feed them as one vector to the encoder. After the latent values are sampled and the reparameterization trick is applied, the latent variable $\boldsymbol{z}$ is concatenated with $\boldsymbol{c}$ again and the vector is fed to the decoder. Fig. 7.1, illustrates the structure and computational flow of the CVAE. After training, when we want to generate an output with label $\boldsymbol{c}$, we sample a $\boldsymbol{z}$ from the prior as we did with VAE, we concatenate it with $\boldsymbol{c}$, and we pass it through the decoder.

7.3 Methods and materials

In this section we will describe our implementation of the CVAE framework for robotic EBL generation.

7.3.1 Dataset

We begin with the description of the animation set and its preprocessing.

EBL animation set augmented with eye LEDs modality

This study used the same animation set of the 36 robotic animations, which was initially selected and evaluated with the procedures described in Chapter 4. For this instance of the animation set, we kept two modalities: the motion sequences and the eye LEDs colour

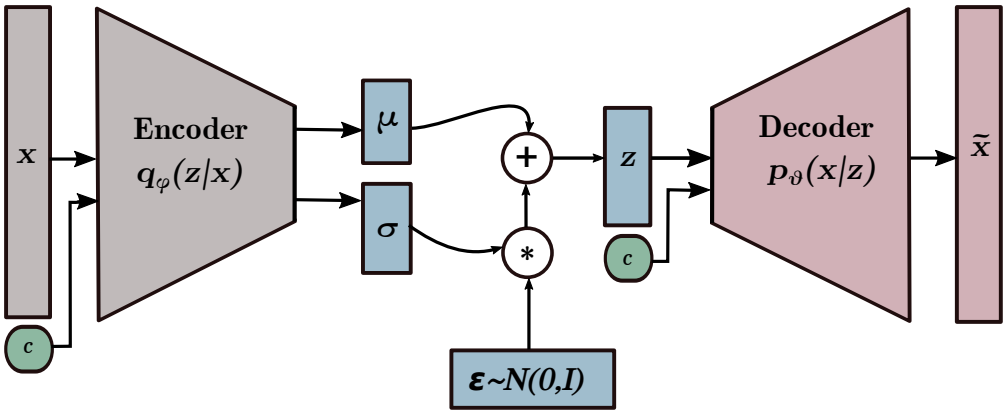

Fig. 7.1 The structure of the CVAE. Compared to the standard VAE, there is an additional input $\boldsymbol{c}$ (green block) that holds a label or an attribute on which we want to condition the generative process. It is concatenated with both input $\boldsymbol{x}$, and the latent $\boldsymbol{z}$.

sequences. The additional modality of the eyes is expected to add more expressiveness to the generated animations. Pepper has two eyes with 8 LEDs each (Fig. 7.2). Each LED is described by 3 values representing an RGB colour. Thus, in total, we have 48 values within a range of $[0, 1]$. As explained in Section 6.2.1, the motion modality comprises keyframes that use 17 values to define the joints' angles of a posture, and the animation is obtained by inbetweening these keyframes with interpolation applied locally at every joint. However, the LEDs representation is different in that their values can be set at any frame along with a duration. Thus, no interpolation is applied. They remain the same until they are manually modified again, or the duration period is over.

Furthermore, we used the valence annotations, collected previously in the study described in Chapter 4. The valence annotations are 36 ratings in the range $[0, 1]$, one for each animation. In the present study, we use them as attributes to condition the CVAE to generate animations of specific valence.

Preprocessing

The preprocessing pipeline is very similar to the one described in Section 6.2.1, with some small modifications to account for the additional LED modality. We had the animations executed by the robot and we recorded both the joints' angles and the LEDs values concurrently. This time we used a sampling rate of 25 frames per second, slightly higher than before to

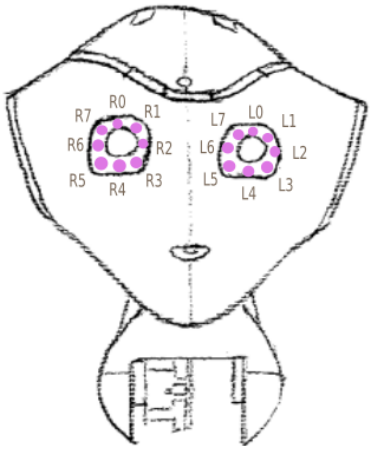

Fig. 7.2 Eye LEDs for Pepper. The robot has 8 LEDs around each eye.

increase our training data given that our input vectors are now larger, with 17 elements for the joints and 48 elements for the LEDs, a total of 65 elements. The StandInit posture was added at the beginning and the end of each animation before the recording to mark a clear onset and offset of the expression.

Next, we doubled the frames by applying the mirroring process described in detail in Section 6.2.1. After this augmentation, we acquired 10148 frames. Regarding the valence attribute, we repeated the valence rating related to each animation for all the frames belonging to it. Therefore, each training example consists of 66 elements: 17 for the joints' angles describing the posture, 48 for the eye LEDs state, and one value indicating the valence rating of the animation to which the frame belongs.

Finally, we rescaled the joints' values in $[0, 1]$ with Min-Max normalization. The scaling was unnecessary for the rest of the values (LEDs or valence rating) since their range was already $[0, 1]$. The dataset was shuffled again, and 80% was used for training, and the rest was preserved for the validation.

7.3.2 The CVAE network implementation

Our CVAE implementation is very similar to the VAE implementation described in Section 6.2.2, with a few alterations to permit the conditioning on the valence attribute, and the extra input data describing the eye LEDs state. We changed the input dimensions from 17 to 66 to include the LEDs state and the valence attribute. The decoder input dimension was also increased by one to match the size of 3D latent $\boldsymbol{z}$ concatenated with the scalar valence attribute. We used MLPs as before but this time we increased the encoder's capacity by one extra fully-connected hidden layer with 512 units. This was decided after several empirical

tests that showed that this configuration was more efficient for the enlarged input and to avoid posterior collapse. Table 7.1 summarizes the architecture characteristics of the CVAE implementation.

Table 7.1 Robotic EBL CVAE architecture specifications

	Encoder	**Decoder**
Input	$\mathbb{R}^{66}$	$\mathbb{R}^{4}$
Reconstruction	NA	$\mathbb{R}^{66}$ (sigmoid)
Hidden layers	4 FC layers	3 FC layers
Hidden units	128, 512, 512, 128	128, 512, 128
Hidden activation	ReLU	
Output activation	Linear	

Training parameters and configuration

All training parameters were the same as in Section 6.2.2, except the mini-batch size and the training epochs. We doubled the mini-batch size from 32 to 64 since we found out empirically that this way the model would converge faster without impact on the loss for the particular configuration. We trained for 250 epochs. The training parameters are summarized in Table 7.2.

7.3.3 Conditional sampling of the latent space

To decide our sampling methodology, we took into consideration two requirements. The first one is to test the hypothesis proposed in Section 6.3.3, according to which, the latent space topology can be exploited to model the arousal dimension of emotion, by controlling the radius of the sampling, which modulates the amplitude and variability of the generated

Table 7.2 Robotic EBL CVAE training parameters

Training parameters
Xavier initialization
Dropout
Adam, $lr = 1e-4$
$\beta = 0.001$
Batch size: 64
Epochs: 250

animation. To sample trajectories with different radius systematically, we can use the spherical grids template from Section 6.2.3, but we need to address a shortcoming discussed in the previous chapter. Recall that when we sampled along the longitudes of a spherical grid, the decoded animations appeared to begin and end abruptly, and always with the same posture. This was because all the longitudes end at the same point, the grid's pole, and in larger grids, this point is far away from the centre. Solving this issue is our second requirement.

To address these two requirements, we decided to change the spherical grid topology to a torus grid. More specifically, we used the topology of a *horn torus* (see Fig. 7.3), which is a torus without a hole, with all the circles forming its tube touching each other at the centre of the 3D latent space where $LD1 = LD2 = LD3 = 0$. In this approach, we sample the latent trajectories along the 8 colourful circles presented in Fig. 7.3. The advantage is that although all the trajectories begin and end at the same point, this point is now the core of the latent space and it decodes into the StandInit posture, the most neutral and symmetric posture. This way, we obtain a clear onset and offset of the animation and we still keep the radius feature with which we will modulate the arousal dimension of the generated animations.

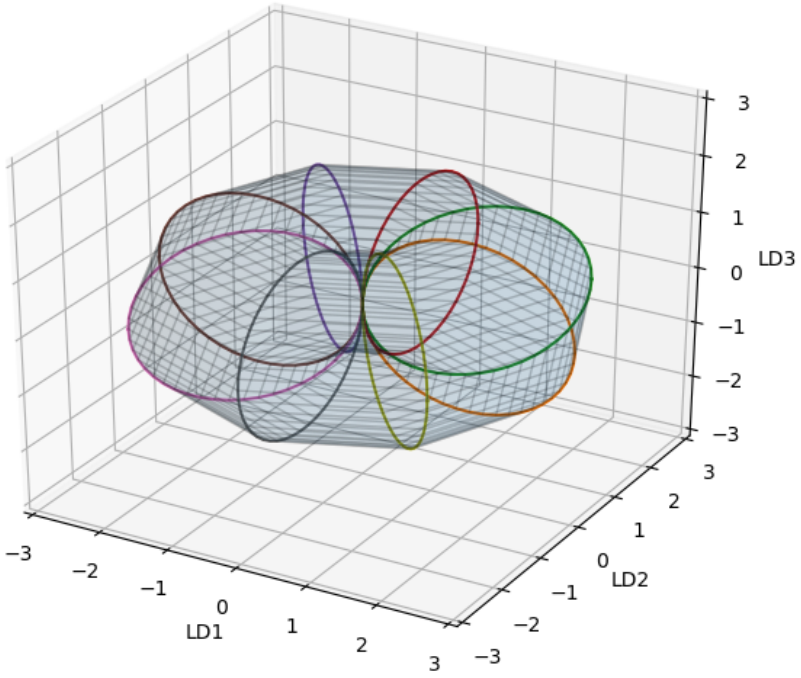

Fig. 7.3 One of the torus grids we used for sampling latent trajectories. The particular one is of radius equal to 3, it has its axis in parallel with the latent dimension LD3, and its center is in the core of the latent space where $LD1 = LD2 = LD3 = 0$. The latent trajectories are sampled along the colored longitude lines.

Arousal conditioning

To condition the generated animations on the arousal as hypothesized to be modulated by the radius of the latent space, we defined three levels of arousal as low, medium, and high, which were obtained by sampling the 3D latent space with a radius of 3, 4, and 5 respectively.

More concretely we sampled 8 circular latent trajectories from torus grids rotated in three ways (so that the central axis is parallel to one of three latent dimensions), and radiuses of 3, 4 and 5 (for example, Fig. 7.3 illustrates 8 circular trajectories from a torus grid of radius equal to 3, and its central axis in parallel with LD3). The radius is defined as the distance from the centre of the latent space to the torus surface's outer point; as if the whole torus is enclosed in a sphere with the same centre. 20 points defined each latent circle. Subsequently, the 20 points of each latent trajectory were interpolated with B-spline interpolation of 15, 20, 25 steps per segment for radius 3, 4, and 5 respectively. We gradually increased the interpolation steps along the segments' size to keep a similar distance between the final points within its trajectory. In total, we derived 72 latent trajectories.

Valence conditioning

To condition the generative process with the valence attribute, we concatenated each of the 72 interpolants with 3 different values representing levels of valence: 0 for negative, 0.5 for neutral, and 1 for positive. Thus we obtain 216 sequences that we can pass through the decoder of the CVAE to get the final animations. Essentially, the valence conditioning is achieved via the CVAE model structure, while the arousal conditioning is accomplished with the sampling method applied to the latent space.

7.3.4 Software

For recording and executing robotic animations with the real Pepper robot we used NAOqi 2.5 SDK[2]. For the robot simulations we used the Choregraphe Suite[3]. The code for the CVAE model was adapted from the TFModelLib collection of neural networks[4]. For the implementation of B-splines we used Splipy[5]. The rest of the code for the CVAE implementation (data preprocessing, sampling, torus grids, etc.) was written in Python 3 with packages such as NumPy [178], SciPy [219], pandas [162], scikit-learn [182], Matplotlib [104].

[2]NAOqi 2.5: http://doc.aldebaran.com/2-5/home_pepper.html
[3]Choregraphe Suite 2.5: http://doc.aldebaran.com/2-5/software/choregraphe/index.html
[4]TFModelLib repository: https://github.com/nhemion/tfmodellib
[5]Splipy v1.3.1: https://sintefmath.github.io/Splipy/

7.4 Conclusion

In this chapter, we presented a Conditional Variational Autoencoder implementation for the generation of multi-modal robotic EBL of targeted valence and arousal for a Pepper robot. In this implementation, we tried to address all the limitations described in Section 6.4.1. Similarly to the previous implementation presented in Chapter 6, the network is trained with a small set of robotic EBL animations, but this time, besides the motion sequences, we included sequences of eye LEDs colours to obtain more expressive animations. Importantly, we aimed to contribute towards generating animations of targeted emotion, by conditioning the generation process with valence labels and a sampling feature that models arousal. Valence labels are used to explicitly condition the CVAE network, while arousal conditioning is achieved by sampling the spherical latent space with different radiuses following our hypothesis formed in Chapter 6. According to this hypothesis, trajectories sampled from near the core of the latent space are decoded into animations of lower activation, and as we move the sampling outwards, the activation becomes higher. Moreover, we modified the sampling strategy by using torus grids instead of spherical grids so that the generated animations exhibit a clear onset and offset with respect to the robot's StandInit posture, which is encoded in the centre of the latent space. A video with the physical robot executing several CVAE generated animations is available online[6].

[6]CVAE generated animation set video: https://youtu.be/wmLT8FARSk0

Chapter 8

Evaluation study of the CVAE model

8.1 Introduction

The objective of the work presented in this chapter is to evaluate the interpretability of the generated animations. We designed and conducted a user study for this purpose. Part of the work presented here is published in [151].

The first research question is whether the valence and arousal ratings given by participants who watched the animations on a physical Pepper robot are affected by the valence and arousal conditioning we used to generate the animations. To answer this question we tested the following two hypotheses:

- **H**1: Valence ratings for animations generated with *negative* valence conditioning will be lower than those of animations generated with *neutral* or *positive* valence conditioning. Valence ratings for animations generated with *neutral* valence conditioning will be lower than those of animations generated with *positive*.

- **H**2: Arousal ratings for animations generated with *low* arousal conditioning will be lower than those of animations generated with *medium* or *high* arousal conditioning. Arousal ratings for animations generated with *medium* arousal conditioning will be lower than those of animations generated with *high*.

The second research question is whether participants will perceive the generated animations' anthropomorphism differently from the original animations. In this aspect, we tested the following hypothesis:

- **H**3: The perception of the robot's anthropomorphism is not significantly different between the *designed* and the *generated* animations.

The chapter is organized as follows: The first section describes the design and implementation of the user study and the statistical methods we used for the data analysis. In the last section, we present and discuss the results obtained from our statistical analyses.

8.2 Methods and materials

In this section we will present the design and implementation of the user study we conducted to evaluate the interpretability of the generated animations, and we will outline the statistical methodology used for the analysis of the collected data.

8.2.1 Experimental design

We begin with the description of the user study design. The study used a physical Pepper robot and aimed to evaluate the interpretability of the CVAE generated animations with respect to the valence and arousal conditioning.

Participants

A total of 20 volunteers were recruited for this study (9 female and 11 male). The mean age was 26.4 years ($SD = 5.36$, $min = 21$, $max = 44$). All participants were employees of SoftBank Robotics, Europe, of various professional roles, e.g., engineers, administrative staff, marketing, etc. They self-reported their experience with the robot's animation capabilities on a scale ranging from 0 (no experience at all) to 10 (extremely familiar). The mean self-reported experience was equal to 6.35 ($SD = 1.81$, $min = 3$, $max = 9$). The experiment was carried out at the company's premises in Paris, France. The experimental protocol was granted ethical approval by the University of Plymouth, Faculty of Science and Engineering Research Ethics Committee.

Regarding the sample size, although it is admittedly small, we considered the high reliability obtained in our first study (Chapter 4), where we followed a very similar design with $N = 20$, and we decided to proceed. In any case, the study we present in this chapter is merely an effort to demonstrate the existence of effects related to the conditioning methodology, and we do not maintain that they necessarily reflect on the broader population. In the next sections, we will examine the results in terms of statistical power and further discuss the sample size.

Procedure

The experimenter briefly explained that this experiment aimed to evaluate a set of body language animations displayed by Pepper. This would require observing the robot placed in front of the participant's seat, approximately 2.5 meters away, and reply to the questions appearing on a screen right in front of the participant. The participants were also told that the entire session would take approximately half an hour, but if at any point they felt tired or unwilling to continue it would be perfectly fine to stop. After the Consent Form was signed, the session would commence. After the session was concluded, the experimenter debriefed the participant by explaining that the collected data would be analyzed to evaluate the presented animations' emotional component. A short unstructured interview was conducted in which participants were asked to comment on their experience.

Stimulus, interface, and questionnaires

The stimuli used for this experiment were animations displayed on the real Pepper robot. We had two kinds of animations: *designed* animations, designed by professional animators with the pose-to-pose method, and *generated* animations, generated with the CVAE and the sampling method we described before.

The session was split into three parts, which we will describe in detail in the following paragraphs. The interface was implemented as a web site running on a local server. The participant would use a play button to give the command for the robot to execute an animation, and afterwards, she would insert her responses in different fields that we will describe in this subsection. In Appendix B, we present screenshots of the interface.

Part A consisted of two trials, in which participants watched the robot executing 9 concatenated *designed* animations, and 9 concatenated *generated* animations. The animations were randomly selected in each group from the respective broader libraries of *designed* and *generated* animations. The two trials appeared in randomized order. After each trial, participants had to fill in a questionnaire comprised of two scales, anthropomorphism and animacy, selected from the Godspeed Questionnaire Series (GQS) [16]. Anthropomorphism is a measure of how humanlike the robot appears to be. The scale comprises five 5-point semantic differentials (Fake/Natural, Machinelike/Humanlike, Unconscious/Conscious, Artificial/Lifelike, and Moving rigidly/Moving elegantly). Animacy is a measure of how lifelike the robot's expression appears. The scale comprises six 5-point semantic differentials (Dead/Alive, Stagnant/Lively, Mechanical/Organic, Artificial/Lifelike, Inert/Interactive, Apathetic/Responsive). This part aimed to compare the *designed* animations to the *generated* ones in terms of the anthropomorphism and animacy degree people attribute to the robot.

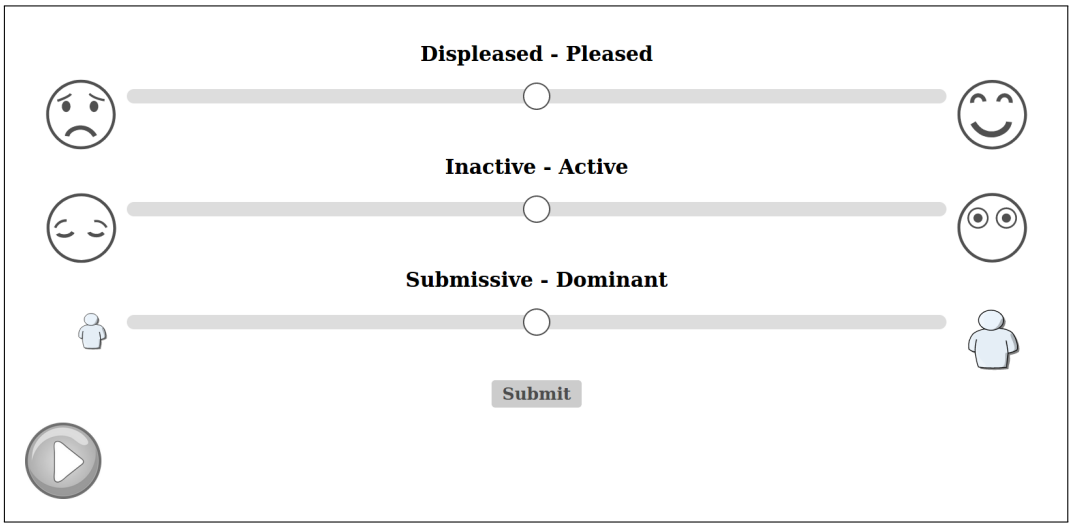

Fig. 8.1 The view we used to collect valence, arousal and dominance ratings. The participant clicks on the play button and watches the real robot executing a generated animation. Subsequently, the user clicks on the three sliders to enter the perceived valence, arousal and dominance scores. The concept is heavily based on the Affective Slider [25].

Part B, used only *generated* animations, since the goal was to have them evaluated in terms of the dimensional emotion interpretation, and more specifically the dimensions of valence, arousal and dominance. This part consisted of 18 randomized trials. In each trial, a single *generated* animation was displayed on the robot when the participant clicked on a digital play button. After the animation was over, the participant would enter her valence, arousal and dominance ratings on three different sliders. This view of the interface is illustrated with a screenshot in Fig. 8.1. It can be seen in the figure that the sliders are titled with the two extremes of each measured emotion dimension: for valence the title is *Displeased - Pleased*, for arousal *Inactive - Active*, and for dominance *Submissive - Dominant*. We selected this verbal description as more intuitive to the participants.

The interface uses visual aids to illustrate the extremes more intuitively. This view's whole idea is based on the Affective Slider [25], a digital scale for the self-assessment of emotion. In our first study, presented in Chapter 4, we used the Affective Slider according to the authors' guidelines; however, in the current study, we have adapted it with several modifications, which we will outline below. The first modification is the addition of a third slider to measure dominance, while the Affective Slider had only two sliders, for valence and arousal. Second, we removed the intensity cue under each slider, otherwise, the participants would have to scroll up and down the web page, since fitting all three sliders along with their intensity cues was not possible. Third, we slightly modified the emoticons based on feedback we received after our first experiment (Chapter 4). The original emoticon related to

the *Pleased* extreme of valence had a very wide open mouth smile, which could confound the valence measure with a heightened arousal interpretation. We wanted to adapt the smile in a more minimal shape that would be perceived as positive, but not as excited, so we used the inverted mouth line of the *Displeased* emoticon. Also, we had the mouth expressions in the two *Inactive - Active* emoticons completely removed, for similar reasons, aiming into keeping a minimal representation for arousal without potential valence confounds arising from the shape of the mouth. Finally, we added two icons to represent the extremes of the *Submissive - Dominant* scale, which we picked based on their minimal design. For a comparison of the two versions, please see Fig. 4.1, where the Affective Slider guidelines were faithfully adapted.

Back to the experimental procedure, the participants were told that they can click the play button to watch the animation a second time if they need to. This was deemed necessary in case the participant's attention was distracted. After watching the animation they could click anywhere on the sliders to enter their ratings. Each slider was constructed with a range in $[0, 1]$ and 100 points of discretization. When the participant had finished configuring the sliders according to her impression, she would click on the submit button. This would lead to a second view, in the same trial, which contained two 5-point Likert scales and the play button again.

The objective of the two Likert scales in the second view aimed at collecting data that would enable us to obtain: 1) an exploratory estimation on how effective are the different conditions and levels of the generated animations in inducing participants' attention, and 2) a secondary evaluation (manipulation check) on whether the generated animations are interpreted as emotional concerning the levels of the conditioning. The two declarations on the Likert scales were "The robot's behaviour draws my attention" and "The robot's expression was emotional". The responses for each of the declarations were: Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree. The participant could watch the animation once more if necessary to recall it under the new set of questions. The attention related Likert scale is adopted from [83], a study that proposes nine single-metric dimensions to evaluate the believability of an artificial agent. The authors of this study, propose a Likert scale with the template phrase, "<X>'s behaviour draws my attention." as a quantifiable metric of the *visual impact* dimension of believability. Using this metric isolated by the rest of dimensions, we can no longer measure believability, but we can obtain a quantification of the visual impact. Although the validity of the measure cannot be guaranteed, we believe that the result can be used as an indicative exploratory measure, that remains to be tested further in a future full-fledged believability analysis. We constructed the

second Likert scale to obtain a secondary, more direct and general assessment on whether participants perceived a *generated* animation as an emotional expression.

The 18 *generated* animations used in Part B were selected randomly from the broader library of generated animations, obtained as we described in Section 7.3.3. We aimed to have 2 animations for each combination of valence and arousal levels. The levels of valence were *negative*, *neutral*, and *positive*, and they were obtained by conditioning the CVAE with a valence attribute of 0, 0.5 and 1 respectively, as described in Section 7.3.3. Arousal levels were defined as *low*, *medium*, and *high*, and they were obtained by sampling the 3D latent space with a radius of 3, 4, and 5 respectively.

Part C was the same as Part A in terms of structure, but different groups of *designed* and *generated* animations were used, to minimize the chances of confounds due to the choice of animations. Furthermore, the *generated* animations used in Part A and C, were excluded from the library before choosing animations for Part B.

The instructions given by the experimenters included demonstrations on how to use the digital play button and the sliders. Specifically for Part B, the task was solicited with the question 'How does the robot feel?', as in our first experiment described in Chapter 4. A practice period of 3 trials preceded it so that the participants could get accustomed to using the sliders and the emotion dimension concepts. In general, when deciding on the number of trials and the length of questionnaires, our main concern was to keep participants' workload as low as possible so that they could retain their concentration and motivation. From pilots conducted before our first study, we had concluded that the whole session should not be longer than half an hour.

8.2.2 Methods for the statistical analysis

Valence, arousal and dominance ratings

This analysis aims to explore how the CVAE conditioning and sampling affected the ratings of valence, arousal and dominance assigned by the participants. Our two independent variables are the following: 1) *v_cond* with levels *negative*, *neutral*, and *positive* representing the valence conditioning of the CVAE generative process using an attribute c equal to 0, 0.5, and 1 respectively, and 2) *a_cond* with levels *low*, *medium*, and *high* representing the arousal sampling of the CVAE's latent space using a radius r equal to 3, 4, and 5 respectively. We summarize them in Table 8.1.

We have three dependent variables, *valence*, *arousal* and *dominance*, which are based on the corresponding slider ratings given by the participants. However, since we had multiple responses from a participant on the same level of the independent variables (e.g., a participant

Table 8.1 Independent variables, their levels and CVAE equivalent parameters.

IV	Levels	CVAE parameter
v_cond	negative	$c = 0$
	neutral	$c = 0.5$
	positive	$c = 1$
a_cond	low	$r = 3$
	medium	$r = 4$
	high	$r = 5$

Note: The independent variables (IV) used in the analyses of the valence, arousal and dominance ratings are named v_cond for valence conditioning, and a_cond for arousal conditioning. The second column presents the levels of each independent variable. The third column maps each level to the equivalent CVAE parameter: c stands for the valence conditioning and r for the radius sampling.

rated 6 animations with *v_cond* = *neutral*), we aggregated the ratings for each participant within each level of the two independent variables *v_cond* and *a_cond*. The aggregation was based on the mean within each level.

The main research question is whether the valence and arousal ratings given by the participants are affected by the valence and arousal conditioning we used to generate the animations. Specifically, we had two hypotheses:

- **H**1: Valence ratings for animations generated with *v_cond* = *negative* will be lower than those of animations generated with *v_cond* = *neutral* or *v_cond* = *positive*. Valence ratings for animations generated with *v_cond* = *neutral* will be lower than those of animations generated with *v_cond* = *positive*.

- **H**2: Arousal ratings for animations generated with *a_cond* = *low* will be lower than those of animations generated with *a_cond* = *medium* or *a_cond* = *high*. Arousal ratings for animations generated with *a_cond* = *medium* will be lower than those of animations generated with *a_cond* = *high*.

Furthermore, following an exploratory analysis, we also examined whether *v_cond* had an effect on the arousal or dominance ratings, and similarly, if *a_cond* had an effect on valence or dominance ratings. Since the experimental design was within-subjects, i.e., each participant evaluated animations in each level of the independent variables, we used repeated-measures one-way analysis of variance (ANOVA). The test was decided to be one-way (with a single independent variable at a time) because the data aggregation applied on each independent variable's levels could not be applied concurrently for both independent variables without inflating the number of appearances of each animation and inserting multiple NaNs. In

Table 8.2 Dependent and independent variables used in the ANOVA analyses

DV (*aggregated response*)	IV (*CVAE conditioning*)
valence	v_cond
	a_cond
arousal	v_cond
	a_cond
dominance	v_cond
	a_cond

Note: The dependent variables (DV) are essentially the aggregated ratings over each level of the independent variable (IV) and each participant. Each row defines the DV and IV of each of the six ANOVAs

total, we applied 6 such tests, with the combinations of dependent and independent variables summarized in Table 8.2.

It should be noted that although a multivariate analysis of variance (MANOVA) including all dependent variables could also be a reasonable choice, this was not deemed necessary since the 6 relationships we examine are conceptually distinct, thus we can use separate ANOVAs. Next, for those relationships that were found significant, we conducted post hoc comparisons between the levels of the independent variables with paired samples t-tests. The p values were adjusted with Bonferroni correction for multiple comparisons error.

Before proceeding with the tests, we checked whether the necessary ANOVA assumptions hold. We checked for outliers using boxplot methods. The normality assumption was tested with the Shapiro-Wilk test and visual inspection of the QQ plots within each independent variable level. The assumption of sphericity was checked with Mauchly's test and Greenhouse-Geisser sphericity correction was applied when the assumption was violated.

Comparison of designed and generated animations

This analysis aims to compare *generated* animations to *designed* ones with regard to the perception of anthropomorphism attributed to the robot that executes them. Furthermore, it examines for possible effects arising from a pretest-posttest experimental design—Part A (pre) and Part C (post) as described in Section 8.2.1—, where the scores on anthropomorphism are collected before and after the main session (Part B), to determine if the perception of anthropomorphism is altered after becoming more familiar with the robot's EBL. Finally, the

present analysis tested for gender effects regarding the anthropomorphism attributed to the robot.

Recall from Section 8.2.1 that the participants watched a set of 9 concatenated *designed* animations and another set of 9 concatenated *generated* animations during Part A of the session, and they had to respond on two scales, Anthropomorphism and Animacy, for each set. This process was repeated with different sets of animations during Part C. Our main hypothesis for this analysis is the following:

H3: The perception of the robot's anthropomorphism as measured by the two scales, Anthropomorphism and Animacy, is not significantly different between the *designed* and the *generated* animations, either in the pretest trials (Part A) or the posttest trials (Part C).

Initially, we computed Cronbach's alpha to estimate the internal consistency of each scale's responses as a measure of reliability, based on which we could decide whether we would collapse the items of each scale to one response per participant and per scale using a measure of central tendency. The decision was based on a threshold $\alpha > 0.7$ which is widely accepted to indicate acceptable internal consistency. Given that result, we proceeded with taking the median of each scale's items and each participant, i.e., one score for Anthropomorphism and another one for Animacy for each participant. Since Likert scales provide ordinal data, the mean is not considered a valid parameter to aggregate their items, and similarly, parametric tests cannot be applied due to the normality assumption, thus we used ordered logistic regression [141] to model the effects. For pairwise differences of the estimated marginal means, the p values of the contrasts were corrected with Tukey's multiple comparisons adjustment.

We fitted 3 models for each scale. The first one aimed to predict the Likert scores using the group of the animation set (designed_pre, generated_pre, designed_post, generated_post) as a predictor, and to examine if there are differences between the *generated* and the *designed* animations, either in the pretest phase or the posttest, according to **H3**. For the second model, we used as a predictor only two groups (pre, post), each containing both designed and generated animations, to explore if there are pretest-posttest differences. The third model aimed to examine potential gender differences affecting the scores. Finally, we applied a likelihood ratio test to each model to test if the proportional odds assumption of the ordered logistic regression holds. This is a crucial assumption for ordinal regression analyses, and it asserts that the independent variables have the same effect on the odds irrespective of the splits between each pair of levels of the ordinal outcome variable. The null hypothesis of the test upholds the proportionality of the odds, thus for the assumption to be justifiable, it must not be rejected.

Attention and emotional content

This analysis examines the effect of the CVAE conditioning variables *v_cond* (levels Negative, Neutral, Positive) and *a_cond* (levels Low, Medium, High) on the 5-point Likert scores of *Attention* ("The robot's behaviour draws my attention") and *Emotion* ("The robot's expression was emotional"). For each Likert item, we fit an ordered logistic regression model and checked for pairwise differences with Tukey's adjustment of p values. The proportional odds assumption is checked again with likelihood ratio tests. This analysis has been added in our experimental design

8.2.3 Software

For executing robotic animations with the real Pepper robot we used NAOqi 2.5 SDK[1]. The interface for the user study data collection was written in Python 2 (NAOqi is not migrated to Python 3) with Django v1.11.10[2]. The statistical analysis of the collected data was carried out in R [186, 227, 117, 217, 50, 138, 228, 37].

8.3 Results

8.3.1 Valence, arousal and dominance ratings

Initially, in Table 8.3, we present the descriptive statistics of the aggregated ratings (valence, arousal, and dominance), with respect to the levels of the explanatory variables *v_cond* and *a_cond*. Only two extreme outliers were detected, one in the relationship where *v_cond* is used to predict arousal, and a second one where *a_cond* is used to predict valence. Since these relationships are not substantial for our main hypotheses and we only have an exploratory interest in them, we decided to proceed. All three dependent variables—valence, arousal and dominance—were found normally distributed at each level of *v_cond* as assessed by Shapiro-Wilk's test ($p > 0.05$). The same was true for arousal and dominance with respect to the levels of *a_cond*, but valence was found to violate the assumption for the *a_cond* level *medium* ($p = 0.029$). Again, taking into account that it is not in our main hypothesis to predict valence ratings with the arousal conditioning of the CVAE, this violation of the normality assumption was not of concern. Furthermore, one-way ANOVA is considered a robust test against the normality assumption.

[1]NAOqi 2.5: http://doc.aldebaran.com/2-5/home_pepper.html
[2]Django web framework: https://www.djangoproject.com/

Table 8.3 Summary statistics for valence, arousal and dominance ratings

Aggregated ratings	v_cond level	Mean	SD	a_cond level	Mean	SD
valence	Negative	0.46	0.12	Low	0.47	0.1
	Neutral	0.54	0.09	Medium	0.54	0.11
	Positive	0.57	0.12	High	0.55	0.13
arousal	Negative	0.55	0.12	Low	0.5	0.12
	Neutral	0.55	0.12	Medium	0.55	0.12
	Positive	0.58	0.11	High	0.63	0.14
dominance	Negative	0.4	0.12	Low	0.42	0.11
	Neutral	0.46	0.1	Medium	0.44	0.09
	Positive	0.52	0.13	High	0.52	0.12

Table 8.4 One-way repeated measures ANOVA tests for valence conditioning (*v_cond*)

DV	F(df)	η_g^2	p	VoA
valence	$F(2,38) = 13.5$	0.16	$< 0.001^{***}$	
arousal	$F(2,38) = 1$	0.01	0.38	outliers
dominance	$F(2,38) = 12.6$	0.16	$< 0.001^{***}$	

First, we will discuss the results for the tests in which we used the valence conditioning of the CVAE (*v_cond*), to predict valence, arousal and dominance ratings. The results of the one-way repeated-measures ANOVAs revealed a significant effect of *v_cond* on the aggregated valence ratings ($F(2,38) = 13.5,\ p < 0.001,\ \eta_g^2 = 0.16$) and on the aggregated dominance ratings ($F(2,38) = 12.6,\ p < 0.001,\ \eta_g^2 = 0.16$). No significant effect was detected on arousal. We summarize the results in Table 8.4, along with the violations of assumption (VoA).

With regard to the arousal conditioning of the CVAE (*a_cond*), the one-way repeated-measures ANOVAs detected significant effects on the aggregated valence ratings ($F(2,38) = 6.47,\ p = 0.004,\ \eta_g^2 = 0.09$), the aggregated arousal ratings ($F(1.5,28.42) = 10.19,\ p = 0.001,\ \eta_g^2 = 0.15$, with Greenhouse-Geisser correction for the degrees of freedom due to sphericity assumption violation), and the aggregated dominance ratings ($F(2,38) = 16.159,\ p < 0.0001,\ \eta_g^2 = 0.166$). The results are summarized in Table 8.5, along with the violations of assumption (VoA).

For both valence and dominance, and *v_cond* as a predictor, the generalized eta-square ($\eta_g^2 = 0.16$) implies a Cohens's f equal to 0.44, while for arousal and dominance, and *a_cond*

Table 8.5 One-way repeated measures ANOVA tests for arousal conditioning (*a_cond*)

DV	F(df)	η_g^2	p	VoA
valence	$F(2,38) = 6.47$	0.09	0.004^{**}	outliers, normality
arousal	$F(1.5,28.42) = 10.19$	0.15	0.001^{***}	sphericity
dominance	$F(2,38) = 16.16$	0.17	$< 0.001^{****}$	

Note: DV = Dependent Variable, F(df) = F statistic and degrees of freedom of numerator and denominator, η_g^2 = generalized eta-squared, p = p values, VoA = violation of assumptions. Degrees of freedom and p values are corrected when sphericity assumption is violated.

as a predictor, the generalized eta-square values ($\eta_g^2 = 0.15$ and $\eta_g^2 = 0.17$ respectively) imply a Cohens's $f > 0.42$. All these effect sizes are considered large [52] and the statistical power for our small sample size ($N = 20$) is greater than 0.81.

In terms of pairwise differences between the levels of the independent variables, we conducted post hoc tests excluding the relationships for which ANOVAs did not reveal significant effects (arousal with *v_cond* as a predictor), and also those for which the normality assumption was violated (valence with *a_cond* as a predictor). All post hoc test results are summarized in Table 8.6.

For the effect of *v_cond* on valence ratings (Fig. 8.2A), paired samples t-tests with Bonferroni correction revealed statistically significant differences for Negative-Neutral ($t = -3.65,\ p = 0.005$) and Negative-Positive ($t = -4.47,\ p < 0.001$). Thus, the results partially support the hypothesis **H1**, in that animations generated with negative valence conditioning are assigned with lower valence ratings compared to animations generated with neutral or positive valence conditioning.

For the effect of *a_cond* on the arousal ratings (Fig. 8.2B), paired samples t-tests with Bonferroni correction detected statistically significant differences for Low-High ($t = -3.62,\ p = 0.005$) and Medium-High ($t = -2.88,\ p = 0.029$). For Low-Medium a statistically significant difference was detected ($t = -2.27,\ p = 0.01$), but it did not survive the Bonferroni adjustment. Thus, the results partially support the hypothesis **H2**, in that animations generated with high arousal conditioning are assigned with higher arousal ratings compared to animations generated with low or medium arousal conditioning.

Furthermore, we conducted post hoc tests to examine the pairwise differences in the levels of *v_cond* and *a_cond* regarding their effect on the aggregated dominance ratings. For *v_cond*, statistically significant differences were detected for Negative-Positive ($t = -4.69,\ p < 0.001$) and Neutral-Positive ($t = -3.46,\ p = 0.008$) (Fig. 8.3A). For *a_cond*, statistically significant differences were detected for Low-High ($t = -4.37,\ p < 0.001$) and

Table 8.6 Post hoc tests for valence, arousal and dominance ratings

DV	IV	Group 1	Group 2	t(df)	p	p (adjusted)
valence	v_cond	Negative	Neutral	$t(19) = -3.65$	0.002	0.005^{**}
		Negative	Positive	$t(19) = -4.47$	< 0.001	$< 0.001^{***}$
		Neutral	Positive	$t(19) = -1.63$	0.12	0.36
dominance	v_cond	Negative	Neutral	$t(19) = -2.06$	0.05	0.16
		Negative	Positive	$t(19) = -4.69$	< 0.001	$< 0.001^{***}$
		Neutral	Positive	$t(19) = -3.46$	0.003	0.008^{**}
arousal	a_cond	Low	Medium	$t(19) = -2.27$	0.04	0.11
		Low	High	$t(19) = -3.62$	0.002	0.005^{**}
		Medium	High	$t(19) = -2.88$	0.01	0.029^{*}
dominance	a_cond	Low	Medium	$t(19) = -0.89$	0.381	1
		Low	High	$t(19) = -4.37$	< 0.001	$< 0.001^{***}$
		Medium	High	$t(19) = -5.24$	< 0.001	$< 0.001^{***}$

Note: DV = Dependent Variable, IV = Independent Variable, t(df) = t statistic and degrees of freedom, p = p values, p (adjusted) = p values with Bonferroni correction.

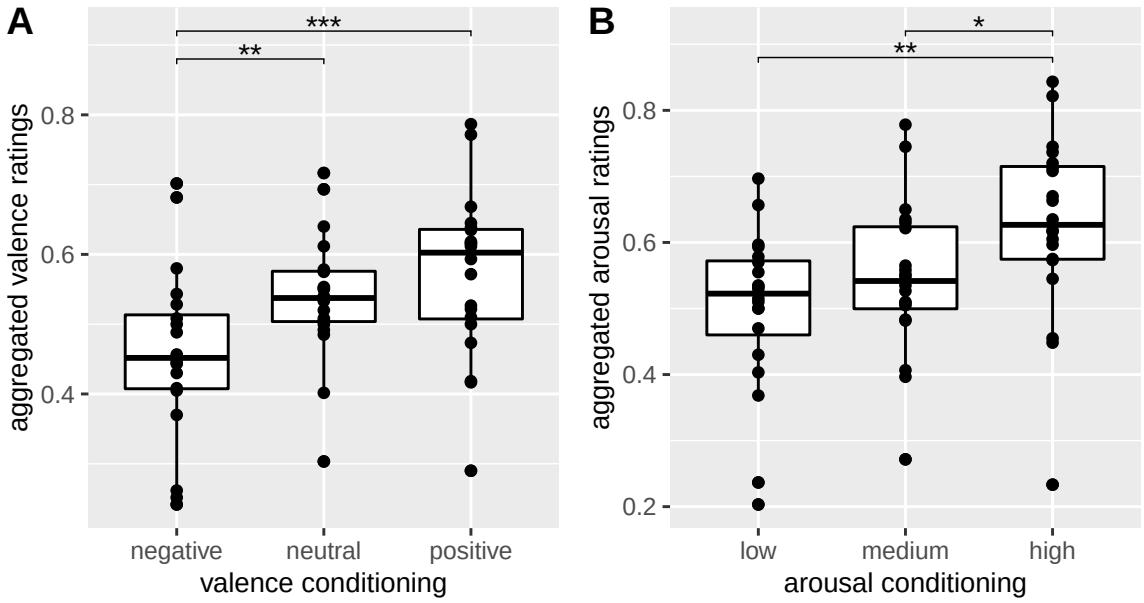

Fig. 8.2 Post hoc t-tests with Bonferroni adjustment for multiple comparisons. A) Differences among the valence conditioning levels with respect to the participants' aggregated ratings of valence. B) Differences among the arousal conditioning levels with respect to the participants' aggregated ratings of arousal. The graph was originally published in [151]

.

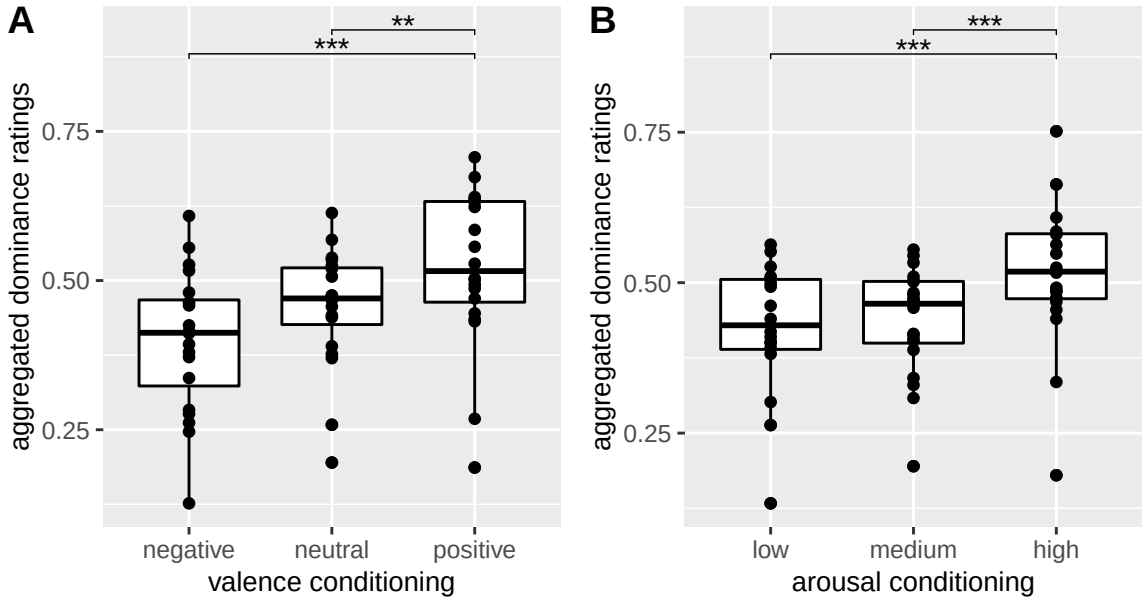

Fig. 8.3 Post hoc t-tests with Bonferroni adjustment for multiple comparisons. A) Differences among the valence conditioning levels with respect to the participants' aggregated ratings of dominance. B) Differences among the arousal conditioning levels with respect to the participants' aggregated ratings of dominance.

Medium-High ($t = -5.24$, $p < 0.001$) (Fig. 8.3B). These results suggest that the CVAE conditioning with *v_cond* and *a_cond* can potentially influence how people perceive the level of dominance in the robot's expression, with positive valence or high arousal conditioning making the robot to appear more dominant in terms of the simulated emotion.

8.3.2 Comparison of designed and generated animations

We begin the presentation of the results with the descriptive statistics presented as bar plots in Fig. 8.4 and as heat maps in Fig. 8.5 for the Anthropomorphism and Animacy scores. The responses are from 1 to 5 (with 5 indicating that the participant attributes a higher degree of Anthropomorphism or Animacy to the robot), and the percentages represent the portion of participants in each response. The basic contrast is *designed* vs *generated* animations presented in the pretest phase (Part A of the session) and then again in the posttest phase (Part C).

For each scale (Anthropomorphism and Animacy), the internal consistency was tested with Cronbach's alpha and was found above our chosen threshold $\alpha > 0.7$ (Table 8.7). Thus, we proceeded with collapsing each scale's semantic differentials (5 items for Anthropomor-

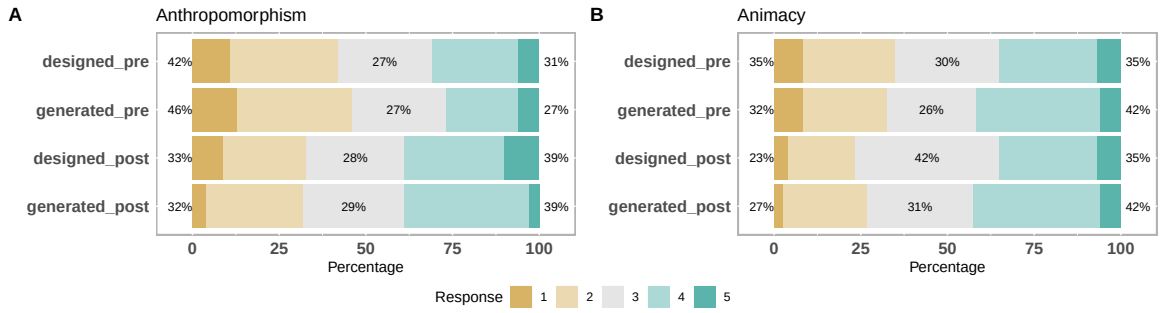

Fig. 8.4 Anthropomorphism (A) and Animacy (B) bar plots for designed vs generated animations in pretest and posttest phase.

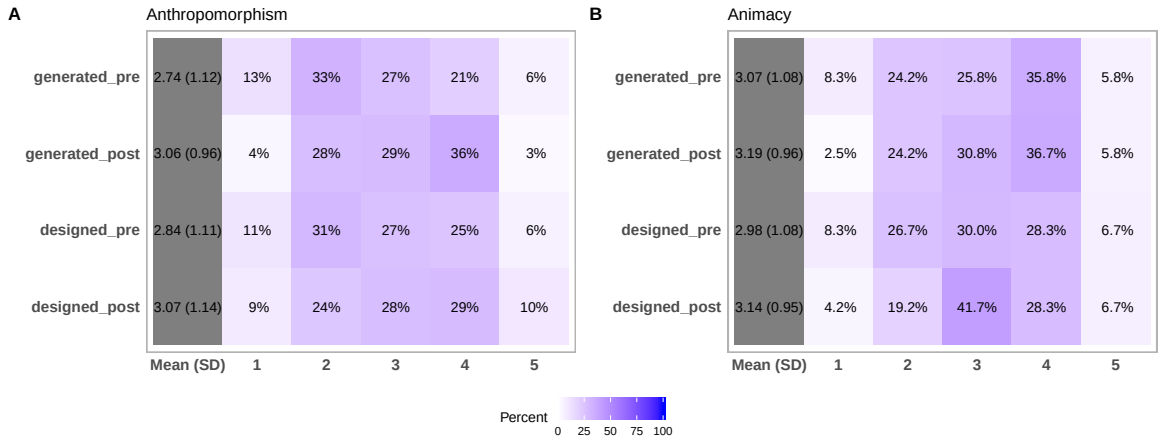

Fig. 8.5 Anthropomorphism (A) and Animacy (B) heat maps for designed vs generated animations in pretest and posttest phase.

phism and 6 for Animacy) in one score per participant and per group, by taking the median value.

Table 8.7 Cronbach's alpha for Anthropomorphism and Animacy scales

	Animacy	Anthropomorphism
designed_pre	0.92	0.89
generated_pre	0.86	0.88
designed_post	0.92	0.92
generated_pre	0.85	0.87

In terms of the hypothesis **H3** testing with ordered logistic regression, none of the groups (designed_pre, generated_pre, designed_post, generated_post) was found to have a statistically significant effect for Anthropomorphism or Animacy (results are summarized in Table 8.8). This result supports our hypothesis that participants do not attribute different Anthropomorphism or Animacy levels to the *designed* animations compared to the *generated* ones, either in the pretest or posttest phase. In Fig. 8.6, we plot the probabilities derived from the predictions of the two ordered logit models of Anthropomorphism and Animacy scores for each group of animations.

Table 8.8 Ordered logistic regression results for Anthropomorphism and Animacy

DV	IV	Coef.	SE	t	p	95% CI
Anthropomorphism	generated_pre	-0.20	0.58	-0.35	0.730	(-1.34, 0.93)
	designed_post	0.57	0.58	0.98	0.328	(-0.57, 1.73)
	generated_post	0.46	0.57	0.81	0.420	(-0.66, 1.60)
Animacy	generated_pre	0.38	0.67	0.57	0.572	(-0.94, 1.71)
	designed_post	1.15	0.66	1.75	0.080	(-0.13, 2.46)
	generated_post	0.81	0.69	1.18	0.240	-0.54, 2.19)
Anthropomorphism	posttest	0.61	0.41	1.49	0.135	(-0.19, 1.42)
Animacy	posttest	0.83	0.49	1.69	0.091	(-0.12, 1.81)
Anthropomorphism	Male	-0.86	0.42	-2.06	0.039*	(-1.69, -0.05)
Animacy	Male	-0.45	0.49	-0.92	0.359	(-1.42, 0.5)

Note: DV = Dependent Variable, IV = Independent Variable, Coef. = Coefficient of ordered logistic regression model, SE = Standard Error, t = t statistic, p = p value, CI = Confidence Interval.

The two models of ordered logistic regression for Anthropomorphism and Animacy with all the pretest and all the posttest animations as a two-level predictor revealed some trends

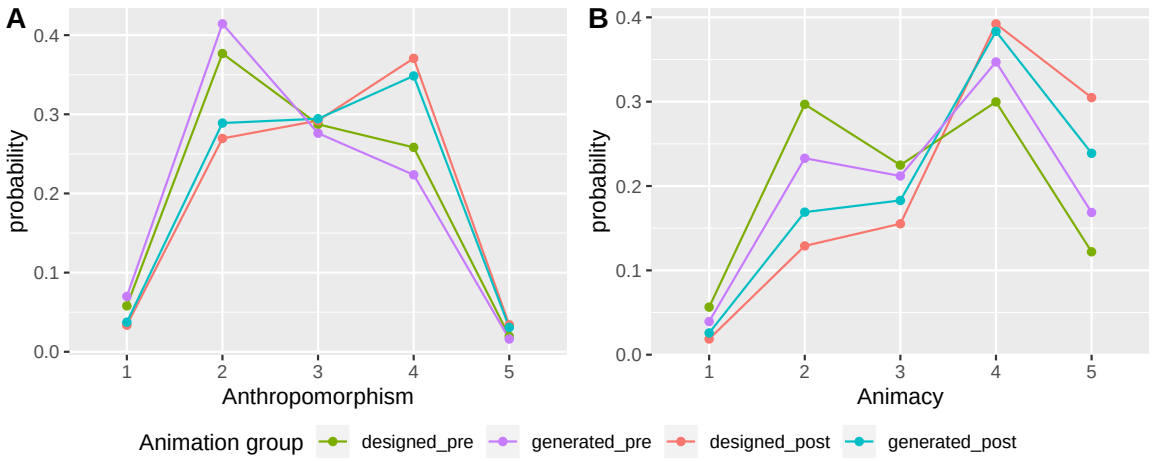

Fig. 8.6 Anthropomorphism and Animacy scores for designed vs generated animations. The *pre* and *post* suffixes indicate the evaluation that was conducted during Part A and Part C of the experimental session respectively. No pairwise significant differences were detected.

for Animacy (Fig. 8.7B), but no significant effects were detected for either scale, suggesting that the interval in between the two evaluations did not impact significantly the perception of Anthropomorphism and Animacy. The detailed results are presented in Table 8.8.

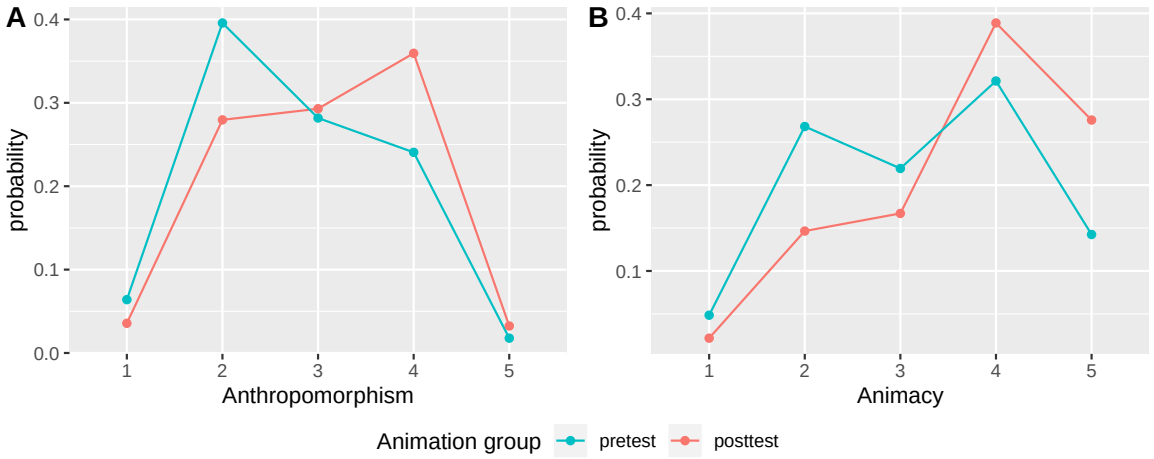

Fig. 8.7 Anthropomorphism and Animacy scores for all the animations evaluated in Part A and Part C. A slight increase in the posttest phase, is not confirmed with statistically significant results in pairwise comparisons.

Finally, regarding the two models that tested for gender effects, female participants gave higher scores than male ($p = 0.039$) on the Anthropomorphism scale, but no statistically significant differences were detected for Animacy ($p = 0.35$). The predictions' probabilities are plotted in Fig. 8.8 and the detailed results are included in Table 8.8.

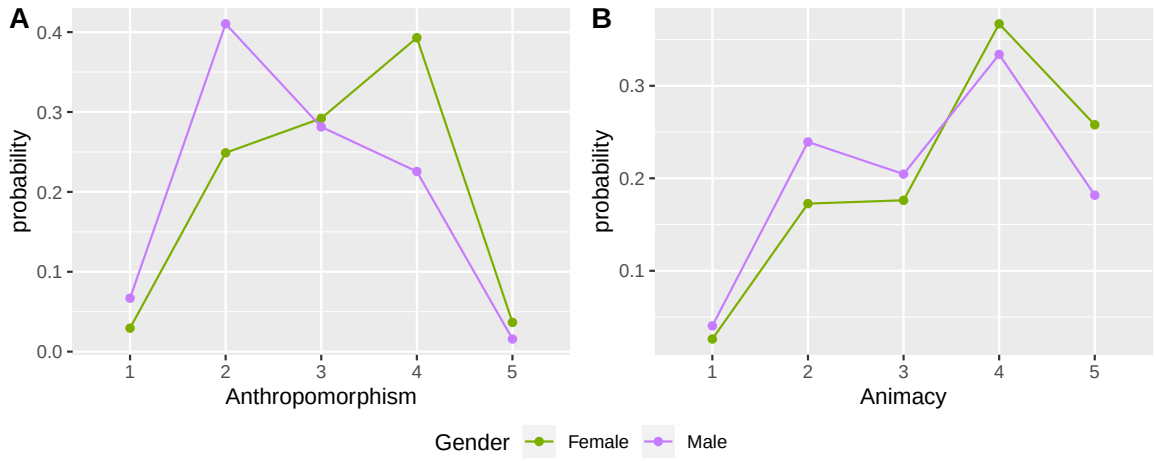

Fig. 8.8 Anthropomorphism and Animacy scores between male and female participants. Female participants gave significantly higher Anthropomorphism ratings.

The likelihood ratio tests checking for violations of the proportional odds assumption did not obtain statistically significant p values for any of the six models (Table 8.9), thus the assumption was considered tenable.

Table 8.9 Proportional odds assumption tests for Anthropomorphism and Animacy

DV	IV	LRT	$p(>\chi^2)$
Anthropomorphism	designed vs generated	11.49	0.24
	pretest vs posttest	2.91	0.41
	gender	1.94	0.58
Animacy	designed vs generated	7.59	0.58
	pretest vs posttest	1.93	0.59
	gender	6.75	0.08

Note: DV = Dependent Variable, IV = Independent Variable, LRT = Likelihood Ratio Test, $p(>\chi^2)$ = p value based on the asymptotic chi-square distribution of the likelihood ratio statistic under the null hypothesis

8.3.3 Attention and emotional content

The descriptive statistics of the Attention and Emotion Likert scales with respect to the valence and arousal conditioning levels are presented in Fig. 8.9 as bar plots and Fig. 8.10 as heat maps with the exact percentage for each response. The two ordered logistic regression models that predicted the Attention Likert scores ("The robot's behaviour draws my atten-

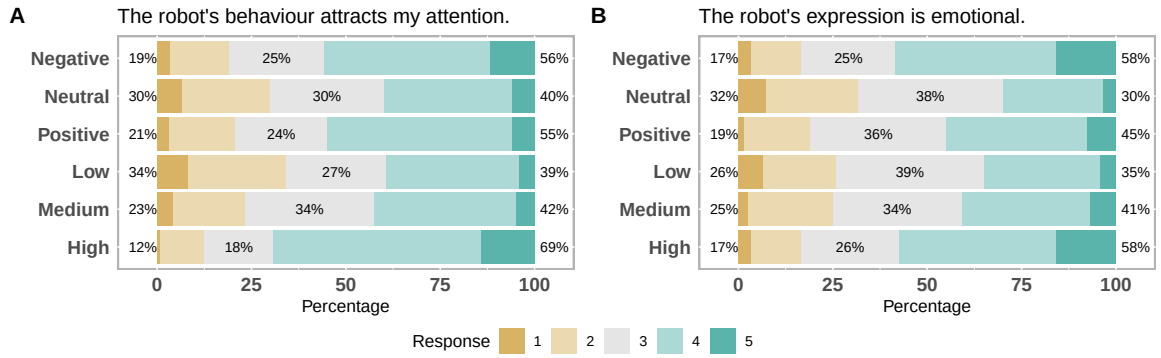

Fig. 8.9 Bar plots for Attention and Emotion Likert scores with respect to the valence (Negative, Neutral, Positive) and arousal (Low, Medium, High) conditioning levels.

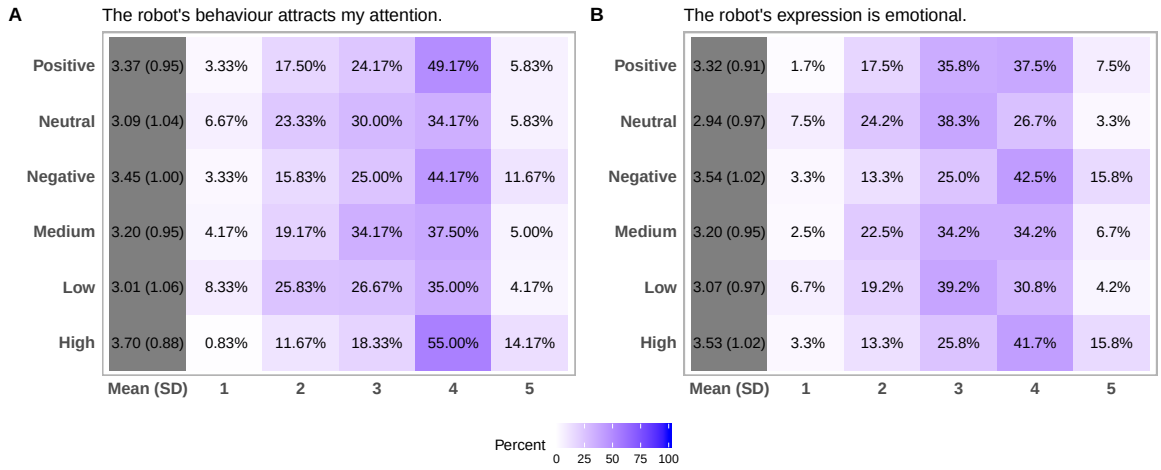

Fig. 8.10 Heat maps for Attention and Emotion Likert scores with respect to the valence and arousal conditioning levels.

tion") with the valence and arousal conditioning of the CVAE as a predictor respectively, detected statistically significant effects. The results are summarised in Table 8.10.

Post hoc tests on the levels of the valence conditioning (*v_cond*) determined that the contrast Negative-Neutral was statistically significant ($z = 2.77,\ p = 0.02$), suggesting that animations generated with negative valence conditioning might attract more attention from participants compared to animations generated with neutral valence. This result is illustrated in Fig. 8.11A, where we present the probabilities of predictions obtained by the model. In the same figure, we observe that the same appears to be true for animations generated with positive conditioning, however, this difference (Neutral-Positive) did not survive the Tukey adjustment of the p value. The results are summarized in Table 8.11.

Post hoc tests with Tukey's p value adjustment for the arousal conditioning (*a_cond*) revealed statistically significant differences for the contrasts Low-High ($z = -5.33,\ p <$

Table 8.10 Ordered logistic regression results for Attention and Emotion

DV	IV	Coef.	SE	t	p	95% CI
Attention	v_cond:Neutral	-0.11	0.17	-0.64	0.52	(-0.44, 0.22)
	v_cond:Negative	0.48	0.17	2.85	$< 0.001^{***}$	(0.15, 0.81)
Attention	a_cond:Medium	0.95	0.18	5.33	$< 0.001^{***}$	(0.60, 1.30)
	a_cond:High	0.29	0.17	1.73	0.08	(-0.04, 0.62)
Emotion	v_cond:Neutral	-0.36	0.17	-2.10	0.04^{*}	(-0.69, -0.02)
	v_cond:Negative	0.74	0.17	4.39	$< 0.001^{***}$	(0.41, 1.07)
Emotion	a_cond:Medium	0.63	0.17	3.71	$< 0.001^{***}$	(0.30, 0.97)
	a_cond:High	0.19	0.17	1.18	0.24	(-0.13, 0.52)

Note: DV = Dependent Variable, IV = Independent Variable, Coef. = Coefficient of ordered logistic regression model, SE = Standard Error, t = t statistic, p = p value, CI = Confidence Interval.

Table 8.11 Post hoc tests for Attention and Emotion

DV	IV	Group 1	Group 2	z	p (adjusted)
Attention	v_cond	Negative	Neutral	2.77	0.02^{*}
		Negative	Positive	0.64	0.8
		Neutral	Positive	-2.16	0.08
Attention	a_cond	Low	Medium	-1.34	0.37
		Low	High	-5.33	$< .001^{***}$
		Medium	High	-4.17	$< .001^{***}$
Emotion	v_cond	Negative	Neutral	4.75	$< .001^{***}$
		Negative	Positive	2.1	0.09
		Neutral	Positive	-2.81	0.01^{**}
Emotion	a_cond	Low	Medium	-0.89	0.64
		Low	High	-3.71	$< .001^{***}$
		Medium	High	-2.85	0.01^{**}

Note: DV = Dependent Variable, IV = Independent Variable, z = z statistic, p (adjusted) = p value with Tukey correction

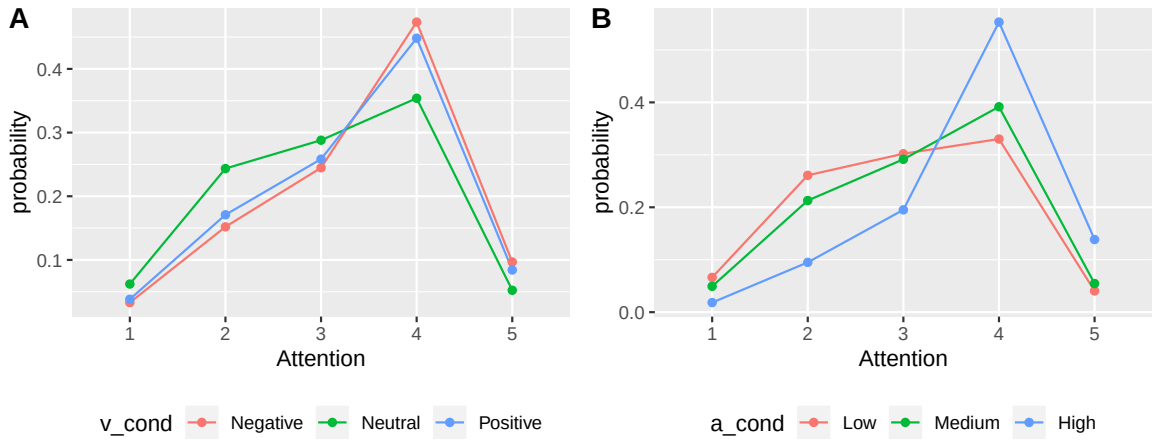

Fig. 8.11 Attention scores ("The robot's behavior draws my attention") for different levels of valence (A) and arousal (B) conditioning. A) Pairwise comparisons revealed significantly higher scores for negative conditioning compared to neutral. B) Pairwise comparisons revealed significantly higher scores for high arousal conditioning compared to both medium and low.

0.001) and Medium-High ($z = -4.17$, $p < 0.001$). The result suggests that potentially animations sampled with larger radius from the latent space tend to draw participants' attention more. The differences are presented in the model's predictions in Fig. 8.11B and the results are summarized in Table 8.11.

The two ordered logistic regression models predicting the Emotion Likert scores ("The robot's expression is emotional") with the valence and arousal conditioning of the CVAE as a predictor respectively, detected statistically significant effects. The results are summarised in Table 8.10. Post hoc tests with Tukey's p value adjustment on the levels of the valence conditioning (*v_cond*) obtained statistically significant results for the contrasts Negative-Neutral ($z = 4.75$, $p < 0.001$) and Neutral-Positive ($z = -2.81$, $p = 0.01$). The results imply (Fig. 8.12A) that animations generated with negative or positive valence conditioning are perceived more as emotional than those generated with neutral valence conditioning. In the figure, a similar trend can perhaps be noted for Positive vs Negative, but this difference did not remain significant after the correction for multiple comparisons. The summary of the results is presented in Table 8.11.

For *a_cond* (arousal conditioning), statistically significant differences were detected for the contrasts Low-High ($z = -3.71$, $p < 0.001$) and Medium-High ($z = -2.85$, $p = 0.01$). This result suggests that animations sampled with a larger radius are perceived as more emotional. See also Fig. 8.12B, and Table8.11 for a summary of the results.

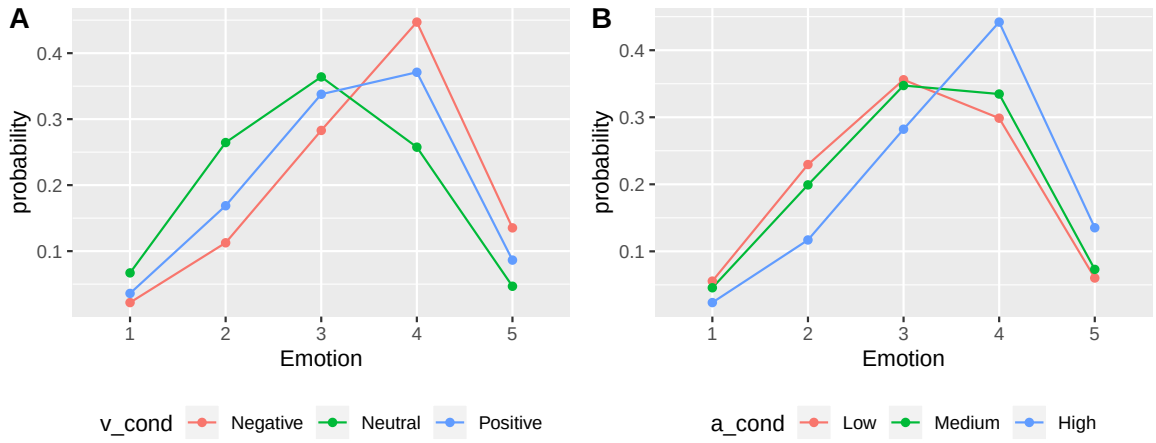

Fig. 8.12 Emotion scores ("The robot's expression is emotional") for different levels of valence (A) and arousal (B) conditioning. A) Pairwise comparisons revealed significantly higher scores for negative and positive conditioning compared to neutral. B) Pairwise comparisons revealed significantly higher scores for high arousal conditioning compared to both medium and low.

Regarding the proportional odds assumption, the likelihood ratio tests of the model terms produced p values greater than 0.05 for all four models, thus the null hypothesis cannot be rejected and the assumption is tenable. The results of the tests are summarized in Table 8.12.

Table 8.12 Proportional odds assumption tests for Attention and Emotion

DV	IV	LRT	$p(>\chi^2)$
Attention	v_cond	2.83	0.83
	a_cond	4.15	0.66
Emotion	v_cond	5.04	0.54
	a_cond	5.67	0.46

Note: DV = Dependent Variable, IV = Independent Variable, LRT = Likelihood Ratio Test, $p(>\chi^2)$ = p value based on the asymptotic chi-square distribution of the likelihood ratio statistic under the null hypothesis

8.4 Conclusion

In this chapter, we presented the implementation of a user study with 20 participants designed to evaluate the interpretability of the generated animations and explore how they compare

with the hand-coded animations in terms of their Anthropomorphism and Animacy attribution. The conclusions are summarized as follows:

1. The participants assigned higher valence ratings to animations that were generated with positive valence conditioning compared to those generated with negative (Table 8.6). The same was true for animations generated with neutral compared to animations generated with negative. However, between positive and neutral the difference was not statistically significant.

2. The participants assigned higher arousal ratings to animations that were generated with high arousal conditioning compared to those generated with medium, or low (Table 8.6). However, the difference was not statistically significant between medium and low.

3. Although we did not implement any conditioning for the dominance dimension of emotion, participants' responses show that animations generated with a positive valence or high arousal, are interpreted as of higher dominance compared to neutral and negative valence or medium and low arousal respectively (Table 8.6).

4. The participants did not perceive differently neither the Anthropomorphism nor the Animacy attributed to the robot between the designed (hand-coded pose-to-pose) animations and the CVAE generated ones.

5. The participants did not perceive differently the Anthropomorphism or the Animacy attributed to the robot at the beginning of the session compared to the end. Some interesting trends show a slight increase though, especially for Animacy (Table 8.8).

6. Female participants assigned higher scores of Anthropomorphism to the robot's EBL (Table 8.8).

7. The CVAE conditioning had some statistically significant effects on the degree participants felt that the robot draws their attention. Negative valence conditioning draws more attention than neutral, and high arousal conditioning draws more attention than low or medium (Table 8.10).

8. Also the CVAE conditioning had some statistically significant effects on how strongly the participants classify robot's EBL as emotion. Negative or positive valence conditioning appears more as an emotional expression compared to neutral. High arousal conditioning is interpreted more as emotional than low or medium (Table 8.10).

In conclusion, the evidence from this study support the idea of using a CVAE to synthesize robotic animations of targeted valence. Also, our hypothesis of using the latent space radius as a sampling feature that modulates the arousal perception of the generated animations appears justifiable by the results. Furthermore, the generated animations do not appear to be less appealing than the designed animations in terms of Anthropomorphism and Animacy. Finally, our exploratory analysis brings some insights regarding the impact of the robot's behaviour on the user's attention (presumably an aspect of believability). Animations of more extreme valence levels, or high arousal, seem to draw more attention, and certainly, this is an intuitive result that supports the impact of the conditioning on users' perception. Similarly, these animations are more strongly classified as emotional. A video with the physical robot executing several CVAE generated animations is available online[3].

8.4.1 Limitations and next steps

Regarding our main hypotheses **H1** and **H2**, our analyses failed to obtain significant p values for the differences between positive and neutral valence conditioning, as well as medium and low arousal conditioning (Fig. 8.2). At first glance, the weak differentiation between neutral-positive levels of valence and low-medium levels of arousal, could be considered as result of the admittedly small sample size ($N = 20$). However, the tests' effect size and statistical power suggest that the sample size of the test is probably adequate.

In an attempt to explain the lack of statistical significance for the particular pairs, we deliberated on the composition of the training set animations. The problem was possibly caused due to an imbalance in the animation set arising from the insufficient representation of positive valence and low arousal animations. Looking back to our first study in Chapter 4, in particular Fig. 4.7 where we present the collected valence and arousal ratings, one notices that participants did not perceive many animations as of very positive valence and very low arousal at the same time. Overall, our framework's inability to generate animations with strong differentiation between neutral and positive valence or low and medium arousal could be explained by the fact that the network was not trained with enough examples from these two categories. Thus, future implementations would be advisable to curate the training set more carefully by collecting more labelled animations of positive valence and low arousal.

Nevertheless, it must be noted that previous studies using a NAO and a Robovie [20, 173] also report limitations in creating EBL expressions that are interpreted as of low arousal and high valence. Hence, it might be the case that this affect subspace is inherently difficult to represent, at least for robots with the particular degrees-of-freedom (≤ 25), or perhaps

[3]CVAE generated animation set video: https://youtu.be/wmLT8FARSk0

emotional states from this affect subspace can be more successfully conveyed through other modalities, such as facial expression.

Chapter 9

Epilogue

9.1 Overall conclusions

We have presented three studies aiming to define a complete methodology for robotic emotional body language (EBL) synthesis. The project was motivated by studying the strengths and weaknesses found in previous work. More specifically, we sought to propose a methodology that addresses three issues in the synthesis of engaging robotic EBL. These drawbacks are the following:

- Hand-coded design (either feature-based or creative) can only produce a limited number of expressions. As a result, the robot's behaviour might appear predictable, repetitive and monotonous, thus, not compelling or engaging enough, especially in long-term HRI.
- The direct use of human EBL as a prototype for robotic EBL design, might not be sufficient to provide the robot with the *illusion of life*, that is, to render it as a believable character. Humans often use subtle ways to express emotions with body language, and they possess more channels of expression and degrees of freedom compared to robots.
- Categorical models of basic emotions, either used in the design phase or the evaluation phase, might be more biased since they collapse granularity and variability in just a few classes of emotions.

To address these issues and develop our alternative methodology, we adopted the following viewpoints:

- Driven by recent advancements in deep learning and particularly the field of generative models, we trained a Variational Autoencoder (VAE) with a small group of robotic

EBL animations to generate numerous new ones. The generated robotic animations appear realistic and smooth. They capture many aspects from the training examples, but they also incorporate a lot of emerging variation. This way, we can equip a robot with countless expressions and ensure that its emotional expression is not repetitive.

- Instead of using human EBL as a prototype, we selected our training examples from a robotic animation library of hand-coded animations, created with the pose-to-pose method used in graphical animation design. This process follows a creative approach, in which the animators conceive expressive postures, directly adapted to the robot morphology. Human EBL is only used as a source of abstract inspiration. Furthermore, this approach is heavily influenced by the *illusion of life* principles used in graphic animation, e.g., exaggeration, to render a character more appealing. This way, we ensure that the robot's emotional expression will be lifelike, believable and engaging.

- We committed to the dimensional model of core affect as the primary emotion representation used in the labelling, the VAE training and sampling, and the evaluation phase. The representation was implemented with valence and arousal dimensions and a range from 0 to 1 with 100 steps of discretization, to ensure that even subtle variations are captured efficiently, both in and the synthesis and the evaluation.

In the first study (Chapter 4), we curated a small set of hand-coded robotic EBL animations created by professional animators for a Pepper robot, and we had twenty participants to watch them on a real Pepper and evaluate them in terms of valence and arousal. We applied a reliability analysis on the collected ratings, and we aggregated them to derive valence and arousal labels. The labelled animations can be used as a training set. In an exploratory analysis of the ratings, we found that valence is more challenging to judge than arousal and that participants tend to judge valence as neutral when they were not confident in their judgement.

In the second study (Chapter 6), we used the motion sequences of the animation set to train a VAE. We studied both the model's latent space and the generated trajectories to understand the model's capacities. The trajectories of the generated animations demonstrated the variation which can be accomplished with the model. Furthermore, we devised a method to systematically sample the spherical latent space of the model with spherical grids. This method enabled us to discern a geometrical feature which can potentially model the arousal dimension of emotion. More specifically, the radius of sampling the spherical latent space affects the amplitude and variability of the generated animations. Small radius, that is, sampling near the core of the latent space, results in generated animations of low

amplitude and variability, while larger radius produces animations of heightened amplitude and variability.

In the third study (Chapters 7 and 8), we aimed to put everything together. The final framework was based on a Conditional Variational Autoencoder (CVAE) to condition the generative process with valence labels collected in the first study. Furthermore, the radius feature for modelling arousal was used for sampling the latent space of the model. We also used the eye LEDs sequences as an additional modality to train the network and generate more expressive animations. We generated a set of 18 animations with three levels of valence conditioning (Negative, Neutral, Positive) and three levels of arousal conditioning (Low, Medium, High).

We conducted a user experiment with 20 participants who watched the generated animations and evaluated them in terms of valence, arousal and dominance. The participants gave higher valence ratings to animations generated with neutral or positive valence conditioning compared to those generated with negative. The participants also assigned higher arousal ratings to animations generated with high arousal conditioning compared to those generated with medium or low. Furthermore, we found that the animation generated with a positive valence or high arousal conditioning were rated as of higher dominance.

We also examined the robot's impact in attracting participants' attention when it performed the generated animations. We used this question as a measure of believability. We found that animations generated with negative valence draw more attention than those generated with neutral valence. Also, animations generated with high arousal conditioning draw more attention than those generated with medium or low. A second question was whether the participants interpreted the generated animations as emotional. Animations generated with negative or positive valence conditioning were found as more emotional. The same was true for animations generated with high arousal conditioning.

The user study also examined how participants evaluated the Anthropomorphism and Animacy of the generated animations compared to designed animations. No significant differences were detected. Regardless of the condition (generated or designed), we found that female participants rated the animations higher in terms of Anthropomorphism compared to male participants.

Taken together, these findings highlight that our proposed methodology is useful in generating animations of recognizable valence and arousal with some limitations that we will discuss in the next section. Furthermore, the results show that the dominance dimension in robotic EBL can be modelled, to some extent, with valence and arousal. Finally, our findings imply that perhaps we can use extreme values for valence conditioning (negative or positive)

or high arousal conditioning to draw user's attention and boost the perception of an emotional response, which appears rather intuitive of course.

9.2 Limitations

Our final CVAE framework did not succeed to generate animations with a clear distinction between neutral and positive valence, or low and medium arousal. We believe that this is related to the training set in which the animations labelled with positive valence and low arousal were underrepresented. This issue was quite striking from our first study. Nevertheless, we proceeded with the data we had to develop a proof of concept methodology and evaluate its effectiveness and its robustness to the limitations arising from the choice of the training set examples. The user study results highlight the importance of a balanced training set to generate animations with distinguishable features in every level of dimensional core affect.

A weakness concerning the final study is that we did not systematically examine the possibility of conditioning the CVAE explicitly with the available arousal labels, as we did with the valence labels. Instead, we modelled arousal by using the radius of sampling the latent space. Although we considered the alternative conditioning with both valence and arousal labels, we decided to proceed with the implicit conditioning using the radius, since it provides a less supervised model. However, a more systematic comparison of the two methods would be useful.

Finally, regarding the results of the final user study, since most participants had a medium to high familiarity with the robot's patterns of motion, the conclusions might conceivably be different for novel users. Ideally, we would have liked to control this factor better and test for such effects too.

9.3 Future directions

Besides addressing the preceding limitations, future work could also extend the framework by implementing the generated animations' dominance conditioning. It would be useful to find a way to condition dominance since that would allow us to distinguish EBL with similar core affect, e.g., fear and anger. Furthermore, adding the modality of sound in training could further increase the animations' expressivity. Audio clips with non-linguistic sounds are already available for the animations used in the training set, and there could be different methods to encompass it, either by concatenating the audio representation with the rest of the

input or by training a completely separate model and use a late fusion of generated motion, eye LEDs and sound.

The above directions aim to improve the EBL synthesis, which is essentially the emotion expression module in the context of the affective loop framework (see Fig. 1.1). However, it is essential to bridge the emotion expression with an emotion synthesis and an emotion adaptation module, both of which will consider the user's affective state. Thus, another critical direction is exploring emotion recognition systems, emotion synthesis, and adaptation architectures that can be used along with the present EBL synthesis to provide robots with adapted and personalized emotional behaviour.

9.4 Applications

In this section, we would like to highlight three applications that used our work. Ruocco et al. [193] used the valence and arousal ratings derived in our first study (Chapter 4) to select animations for a Nao robot used as a distraction to keep anxiety levels low in children during vaccination. They found that the anxiety levels were reduced for children in the two groups treated with the distraction strategy, compared to the third group, in which the robot did not exhibit any mood-based animation. Similarly, our valence and arousal ratings were used by Rossi et al. [190] to select animations for a Pepper robot with the task to give movie recommendations. The study compares the effect of using coherent and incoherent combinations of robotic EBL and movie genres. For example, the drama genre is associated with negative valence and low arousal in the coherent condition, while in the incoherent condition, it is associated with positive valence and high arousal animations. The authors were interested in finding how the robot's behaviour affects participant's emotion, engagement and attention, which were measured automatically from the facial expression with Affectiva's Machine Learning algorithm [160]. The authors found that measures of joy, smiling and engagement were significantly higher for female participants when presented with the incoherent EBL, suggesting that mismatching the context and robot expression might have an amusing and engaging result for females. However, in the case of recommending movies in the Comedy genre, incoherent robot expressions using animations of negative valence had resulted in significantly higher sadness measured by Affectiva's facial expression algorithm, suggesting that in this setup, the participant's mood was potentially overridden by the robot's expression.

Finally, in a proposed application scenario for the APRIL (Applications of Personal Robotics for Interaction and Learning) Innovative Training Network[1], a personalized robot

[1] APRIL ITN: https://www.fose1.plymouth.ac.uk/socem/crns/april/

narrator with adapted emotional body language was developed, integrating work on multimodal open-set person identification by Irfan et al. [107] and our Conditional Variational Autoencoder for generated robotic EBL [151] described in Chapter 7.

More specifically, in this demonstration, Pepper identifies the users, retrieves their profiles (e.g., age, interests) and narrates personalized content with emotional body language adapted to their profiles and the content's valence and arousal. For the emotion adaptation part, the user's age profile and the content's genre was used to select the robotic EBL. The strategy was to use neutral valence and low arousal for adults listening to a science news extract, while for a child user listening to a fairy tale, the entire spectrum of valence and arousal levels were used. In particular, for the fairy tale narration, we were based on previous work from Yves Bestgen [24], in which participants used valence scores to annotate sentences in four texts, among which, the fairy tale we used ("The Little Match Girl" by Hans Christian Andersen). To adapt the robotic EBL to the fairy tale narration, we initially timed each sentence while Pepper narrated it. Then we used the sentence's valence annotation from Bestgen, to select the valence level for our generated expressions. The timing was used to select the radius of sampling. A video of the demonstration is available online[2].

9.5 A note on ethical considerations

A paramount concern related to emotional behaviour synthesis is about the potential ethical implications of deploying robots with persuasive affective personalities. Perhaps these technologies are currently not sophisticated enough to alarm society, but it is crucial to anticipate and examine such implications before technology arrives at a more mature stage. Furthermore, even if the current affective robots are not persuasive enough as emotional agents for a typical adult, the same might not be the case with children, especially young ones, or other vulnerable groups.

Khan et al. [114] explored how ninety children interacted socially with the humanoid robot Robovie. The children were of ages 9, 12 and 15. The study reports that 80% of the children believed that the robot was intelligent, and 60% concluded that it had feelings. The authors also examined the morality attributed to the robot. They had the children play the "I Spy" game with the robot when an experimenter interrupted and announced to Robovie that it had to go into a storage closet. The robot protested saying it was not fair, that it was not given enough chances, that its feelings were hurt, and that the closet was dark and scary. According to 88% of the children, the robot was treated unfairly, while 54% thought it was not right to put the robot in the closet. Also, 81% of the children said they felt like they

[2]Personalized robot narrator with adapted emotional body language: https://youtu.be/fBIIn0PQGSA

wanted to comfort Robovie when the robot is sad. Overall, the children conceptualized the robot as a mental, social, and partly moral entity, although these results were less robust for the children in age 15.

In the same study, the authors distinguish two design stances and discuss the related ethical concerns. In the first, robotic behaviours are designed to appear as having human qualities, such as affect expression and moral standing. However, since robots can not possess such qualities, from a philosophical point of view, this design stance might involve a form of intentional emotional deception [51]. Furthermore, children or vulnerable adults might develop emotional attachment towards the robot companion. This could be a problem if such attachment causes a child to seek less of human companionship [215], or if the adults, in the child's environment, tend to sidestep their responsibilities because the robot companion fulfils the gap [54]. On the other hand, robotic behaviours that do not entail some moral or emotional agency, essentially a robot that allows being treated like a machine or an object accepting commands, might evoke carryover effects between mistreating robots and mistreating people [114].

Older persons comprise another sensitive group we have to consider when designing affective human-robot interaction applications. Robot companions endowed with affect recognition or expression skills could be useful in robot-assisted interventions for older adults suffering from emotional distress or cognitive impairments. In this context, several studies have shown positive effects arising from affective robot companions in therapeutic interventions. For example, dementia patients have shown increased communication scores [212] after interacting with a Sony AIBO robotic dog. In studies using the same robot in long term care facilities, the findings indicate improvement in well-being [115]. Decrease of loneliness and stress have even been reported at levels similar to those accomplished by interacting with real pet-animals [9]. Nevertheless, several ethical issues have been identified. A significant concern is that robot companions might be progressively accepted as adequate replacements for human love and attention, providing carers with a justification to leave an older person alone for longer intervals [202]. Considering potential attachment that can arise from such interactions, another ethical concern is the lack of authenticity in these relationships and the fact that they essentially presuppose a form of deception, encouraging a false perception of the world [209] (see also [202] on the risk of infantilization of older adults).

Affective robot companions designed to evoke emotional attachment by using deception or systematic delusion might comprise a form of unethical technology [29, 39–41, 202, 209]. In contrast to the intentional suspension of disbelief, which enables humans to temporarily accept a false reality to experience emotions and enjoy fictional artefacts, systematic de-

ception that encourages adults with cognitive impairments or children to form emotional relationships with companion robots can be problematic. This ethical perspective is captured in the principles of robotics which were adopted in 2010 by the British Engineering and Physical Sciences Research Council (EPSRC)[3], in an attempt to regulate robot ethics at an early stage of development of this critical technology. More specifically, the fourth principle addresses these ethical concerns in a succinct way:

> **Legal formulation:** Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.
>
> **General audience formulation:** Robots are manufactured artefacts: the illusion of emotions and intent should not be used to exploit vulnerable users.

In the related commentary, the contributors of the principles of robotics [29] acknowledge the tradeoff between enhancing an affective robot's persuasiveness to fulfil the role of a companion, and the risk arising from using this persuasiveness to exploit people who might feel emotionally attached to their robot companions. Transparency seems to be the most promising tool we have to resolve this tradeoff, and in line with it, the commentary concludes that "although it is permissible and even sometimes desirable for a robot to sometimes give the impression of real intelligence, anyone who owns or interacts with a robot should be able to find out what it really is and perhaps what it was really manufactured to do."

Joanna Bryson, a contributor of the EPSRC principles of robotics, has been a long-time advocate for the importance of transparency in designing, manufacturing and advertising companion robots [38–41]. She urges roboticists to use explicit ways to emphasize the mechanical nature of robots and to avoid creating robots that would predispose people to assign a human-like status to them [38]. Bryson warns that deceptive design and marketing obscuring that a robot companion is essentially a machine, can cause vulnerable persons to over-identify with it and perceive it as an animate being or a moral agent [39]. In such cases, even if making a robot suffer is not technically plausible, vulnerable persons might feel the opposite is true, or they might even feel inclined to save a robot from a dangerous situation putting their own lives at risk. Furthermore, persons who perceive a robot as animate might develop an urge to sacrifice valuable resources, money or time, to address possible artificial needs, which in some cases might have been designed purposefully to exploit susceptible humans.

[3]EPSRC Principles of robotics https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

Despite the challenges, Sharkey and Sharkey suggest that it is still feasible to deploy robot companions in older persons' lives ethically by introducing them as social facilitators to induce interactions with other people [202]. An older person could be given a robot and be asked to talk about it, show it to other people, and essentially use it as a medium for human-human interaction.

In conclusion, this note on ethical considerations about the design and the use cases of affective robotics aims to acknowledge the potential social risks and support the necessity of further experiments investigating possible vulnerabilities that can be used to exploit humans. A growing body of interdisciplinary studies gradually provides informative insights on issues of ethics. So far, it seems that even when affective robotic expressions are deemed as necessary in some application scenarios, we can still use transparency in several levels, e.g., design and marketing, to emphasize the mechanical nature of a robot. Also, for companion robots used by vulnerable individuals, psychological supervision could be crucial to determine whether the user intentionally suspends her disbelief to enjoy the illusion of life effect or is getting carried away to perceive the robot as animate. Hopefully, as society gets more familiar with and educated on the true nature and potential of such technologies, people would be less susceptible to personify their robots. Nevertheless, policies such as the fourth principle of EPSRC using legal formulation to create an early framework of guidance for both users and roboticists, are definitely in the right direction. Hopefully, such initiatives will keep involving as knowledge is progressing.

References

[1] Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. (2018). Fixing a broken ELBO. In *International Conference on Machine Learning*, pages 159–168.

[2] Anders, S., Lotze, M., Erb, M., Grodd, W., and Birbaumer, N. (2004). Brain activity underlying emotional valence and arousal: A response-related fMRI study. *Human Brain Mapping*, 23(4):200–209.

[3] Angel-Fernandez, J. M. and Bonarini, A. (2016). Robots showing emotions. *Interaction Studies*, 17(3):408–437.

[4] Arbib, M. A. and Fellous, J.-M. (2004). Emotions: From brain to robot. *Trends in Cognitive Sciences*, 8(12):554 – 561.

[5] Argyle, M. (1975). *Bodily communication*. Methuen, London, UK.

[6] Aviezer, H., Trope, Y., and Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111):1225–1229.

[7] Badler, N., Allbeck, J., Zhao, L., and Byun, M. (2002). Representing and parameterizing agent behaviors. In *Proceedings of Computer Animation 2002 (CA 2002)*, pages 133–143. IEEE.

[8] Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58.

[9] Banks, M. R., Willoughby, L. M., and Banks, W. A. (2008). Animal-assisted therapy and loneliness in nursing homes: use of robotic versus living dogs. *Journal of the American Medical Directors Association*, 9(3):173–177.

[10] Barrett, L. F. (2006a). Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1):28–58.

[11] Barrett, L. F. (2006b). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1):20–46.

[12] Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

[13] Barrett, L. F., Lindquist, K. A., Bliss-Moreau, E., Duncan, S., Gendron, M., Mize, J., and Brennan, L. (2007). Of mice and men: Natural kinds of emotions in the mammalian brain? A response to Panksepp and Izard. *Perspectives on Psychological Science*, 2(3):297–312.

[14] Barrett, L. F., Mesquita, B., and Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290.

[15] Bartels, R. H., Beatty, J. C., and Barsky, B. A. (1987). *An introduction to splines for use in computer graphics and geometric modeling*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[16] Bartneck, C., Croft, E., and Kulic, D. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81.

[17] Bartneck, C. and Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. In *13th IEEE International Workshop on Robot and Human Interactive Communication*, pages 591–594.

[18] Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125.

[19] Beck, A., Cañamero, L., and Bard, K. A. (2010a). Towards an affect space for robots to display emotional body language. In *19th International Symposium in Robot and Human Interactive Communication*, pages 464–469.

[20] Beck, A., Hiolle, A., Mazel, A., and Cañamero, L. (2010b). Interpretation of emotional body language displayed by robots. In *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments*, pages 37–42. ACM.

[21] Beck, A., Stevens, B., Bard, K. A., and Cañamero, L. (2012). Emotional body language displayed by artificial agents. *ACM Transactions on Interactive Intelligent Systems*, 2(1):Article 2.

[22] Becker, C., Kopp, S., and Wachsmuth, I. (2004). Simulating the emotion dynamics of a multimodal conversational agent. In André, E., Dybkjær, L., Minker, W., and Heisterkamp, P., editors, *Affective Dialogue Systems*, pages 154–165, Berlin, Heidelberg. Springer Berlin Heidelberg.

[23] Bengio, Y., Courville, A., and Vincent, P. (2012). Representation learning: A review and new perspectives. *arXiv e-prints*, page arXiv:1206.5538.

[24] Bestgen, Y. (1994). Can emotional valence in stories be determined from words? *Cognition and Emotion*, 8(1):21–36.

[25] Betella, A. and Verschure, P. F. M. J. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLoS ONE*, 11(2):e0148037.

[26] Bethel, C. L. and Murphy, R. R. (2008). Survey of non-facial/non-verbal affective expressions for appearance-constrained robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(1):83–92.

[27] Bevacqua, E., Mancini, M., Niewiadomski, R., and Pelachaud, C. (2007). An expressive ECA showing complex emotions. In *Proceedings of the AISB annual convention, Newcastle, UK*, pages 208–216.

[28] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

[29] Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorell, T., Wallis, M., Whitby, B., and Winfield, A. (2017). Principles of robotics: regulating robots in the real world. *Connection Science*, 29:124–129.

[30] Borod, J. C., Cicero, B. A., Obler, L. K., Welkowitz, J., Erhan, H. M., Santschi, C., Grunwald, I. S., Agosti, R. M., and Whalen, J. R. (1998). Right hemisphere emotional perception: evidence across multiple channels. *Neuropsychology*, 12(3):446.

[31] Borod, J. C. and Koff, E. (1984). Asymmetries in affective facial expression: Behavior and anatomy. In Fox, N. A. and Davidson, R. J., editors, *The Psychobiology of Affective Development*, chapter 8, pages 293–321. Psychology Press, London, 1st edition.

[32] Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59(4):291–294.

[33] Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015). Generating sentences from a continuous space. *arXiv e-prints*, page arXiv:1511.06349.

[34] Breazeal, C. (2009). Role of expressive behaviour for robots that learn from people. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3527–3538.

[35] Breazeal, C., Dautenhahn, K., and Kanda, T. (2016). Social robotics. In Siciliano, B. and Khatib, O., editors, *Springer Handbook of Robotics*, Springer Handbooks, pages 1935–1972. Springer International Publishing, Cham.

[36] Broekens, J. (2012). In defense of dominance: PAD usage in computational representations of affect. *International Journal of Synthetic Emotions*, 3(1):33–42.

[37] Bryer, J. and Speerschneider, K. (2016). *likert: Analysis and visualization Likert items*. R package version 1.3.5. https://CRAN.R-project.org/package=likert.

[38] Bryson, J. J. (2000). A proposal for the humanoid agent-builders league (HAL). In *AISB'00 Symposium on Artificial Intelligence, Ethics and (Quasi-) Human Rights*, pages 1–6.

[39] Bryson, J. J. (2010). Robots should be slaves. In Wilks, Y., editor, *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, pages 63–74. John Benjamins Publishing.

[40] Bryson, J. J. (2017). The meaning of the EPSRC principles of robotics. *Connection Science*, 29(2):130–136.

[41] Bryson, J. J. and Kime, P. P. (2011). Just an artifact: why machines are perceived as moral agents. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pages 1641–1646.

[42] Bull, P. E. (1987). *Posture and gesture*. Pergamon Press, Oxford, UK.

[43] Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv e-prints*, page arXiv:1804.03599.

[44] Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., and Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, page 520–527, New York, NY, USA. Association for Computing Machinery.

[45] Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., and Yan, H. (2001). More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14(1):55 – 64.

[46] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer graphics and interactive techniques*, pages 413–420.

[47] Cassell, J., Pelachaud, C., Badler, N. I., Steedman, M., and Achorn, B. (2000). *Embodied conversational agents*. MIT Press, Cambridge, Massachusetts.

[48] Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Variational lossy autoencoder. *arXiv e-prints*, page arXiv:1611.02731.

[49] Chi, D., Costa, M., Zhao, L., and Badler, N. (2000). The EMOTE model for effort and shape. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, page 173–182, USA. ACM Press/Addison-Wesley Publishing Co.

[50] Christensen, R. H. B. (2019). *ordinal: Regression models for ordinal data*. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

[51] Coeckelbergh, M. (2012). Are emotional robots deceptive? *IEEE Transactions on Affective Computing*, 3(4):388–393.

[52] Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, 1(3):98–101.

[53] Coulson, M. (2004-06-01). Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139.

[54] Cowie, R. (2014). Ethical issues in affective computing. In Calvo, R., D'Mello, S., Gratch, J., and Kappas, A., editors, *The Oxford Handbook of Affective Computing*, chapter 24, pages 334–348. Oxford University Press, Inc., USA, 1st edition.

[55] Cowie, R., McKeown, G., and Douglas-Cowie, E. (2012). Tracing emotion: An overview. *International Journal of Synthetic Emotions*, 3(1):1–17.

[56] Dael, N., Mortillaro, M., and Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion*, 12(5):1085–1101.

[57] Damiano, L. and Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology*, 9:468.

[58] Darwin, C. (1872). *The expression of the emotions in man and animals*. John Murray, London, England.

[59] Davidson, R. J. (1992). Anterior cerebral asymmetry and the nature of emotion. *Brain and cognition*, 20(1):125–151.

[60] Davidson, R. J. and Irwin, W. (1999-01). The functional neuroanatomy of emotion and affective style. *Trends in Cognitive Sciences*, 3(1):11–21.

[61] de Gelder, B., Van den Stock, J., Meeren, H. K., Sinke, C. B., Kret, M. E., and Tamietto, M. (2010). Standing up for the body. Recent progress in uncovering the networks involved in the perception of bodies and bodily expressions. *Neuroscience & Biobehavioral Reviews*, 34(4):513 – 527.

[62] Destephe, M., Henning, A., Zecca, M., Hashimoto, K., and Takanishi, A. (2013). Perception of emotion and emotional intensity in humanoid robots gait. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1276–1281.

[63] Diedenhofen, B. (2016). *cocron: Statistical comparisons of two or more alpha coefficients*. R package version 1.0-1. http://comparingcronbachalphas.org.

[64] Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv e-prints*, page arXiv:1606.05908.

[65] Ekman, P. (1971). Universals and cultural differences in facial expressions of emotions. In *Nebraska Symposium on Motivation*, volume 19, pages 207–283.

[66] Ekman, P. and Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review*, 3(4):364–370.

[67] Ekman, P. and Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of personality and Social Psychology*, 29(3):288.

[68] Ekman, P. and Friesen, W. V. (1978). Facial action coding system: A technique for the measurement of facial movement. *Palo Alto, California: Consulting Psychologists Press*.

[69] Embgen, S., Luber, M., Becker-Asano, C., Ragni, M., Evers, V., and Arras, K. O. (2012). Robot-specific social cues in emotional body language. In *IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 1019–1025.

[70] Erden, M. S. (2013). Emotional postures for the humanoid-robot Nao. *International Journal of Social Robotics*, 5(4):441–456.

[71] Fast, J. (1970). *Body Language*. Henry Holt and Company.

[72] Feldt, L. S., Woodruff, D. J., and Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, 11(1):93–103.

[73] Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42:143–166.

[74] Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057.

[75] Foster, M. E. (2007a). Assocating facial displays with syntactic constituents for generation. In *Proceedings of the Linguistic Annotation Workshop*, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics.

[76] Foster, M. E. (2007b). Enhancing human-computer interaction with embodied conversational agents. In Stephanidis, C., editor, *Universal Access in Human-Computer Interaction. Ambient Interaction*, pages 828–837, Berlin, Heidelberg. Springer Berlin Heidelberg.

[77] Foster, M. E. (2019). Face-to-face conversation: Why embodiment matters for conversational user interfaces. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19, New York, NY, USA. Association for Computing Machinery.

[78] Foster, M. E. and Oberlander, J. (2007). Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, 3(41):305–323.

[79] Frank Thomas, O. J. (1995). *The illusion of life: Disney animation*. Disney Editions.

[80] Gamer, M., Lemon, J., and Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement*. R package version 0.84.1. https://CRAN.R-project.org/package=irr.

[81] Gendron, M., Roberson, D., van der Vyver, J. M., and Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, 14(2):251.

[82] Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256.

[83] Gomes, P., Paiva, A., Martinho, C., and Jhala, A. (2013). Metrics for character believability in interactive narrative. In *Proceedings of the 6th International Conference on Interactive Storytelling - Volume 8230*, ICIDS 2013, page 223–228, Berlin, Heidelberg. Springer-Verlag.

[84] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. The MIT Press.

[85] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. *arXiv e-prints*, page arXiv:1406.2661.

[86] Gratch, J. and Marsella, S. (2001). Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. In *Proceedings of the fifth international conference on Autonomous agents*, pages 278–285.

[87] Gratch, J. and Marsella, S. C. (2015). Appraisal models. In Calvo, R., D'Mello, S., Gratch, J., and Kappas, A., editors, *The Oxford Handbook of Affective Computing*, chapter 5, pages 54–67. Oxford University Press, Inc., USA, 1st edition.

[88] Gunes, H. and Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99.

[89] Gunes, H. and Piccardi, M. (2007). Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30:1334–1345.

[90] Gunes, H., Schuller, B., Pantic, M., and Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *Face and Gesture*, pages 827–834.

[91] Haarbach, A., Birdal, T., and Ilic, S. (2018). Survey of higher order rigid body motion interpolation methods for keyframe animation and continuous-time trajectory estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 381–389.

[92] Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23–34.

[93] Harmon-Jones, E., Harmon-Jones, C., and Summerell, E. (2017). On the importance of both dimensional and discrete models of emotion. *Behavioral Sciences*, 7(4):66.

[94] Hartmann, B., Mancini, M., and Pelachaud, C. (2006). Implementing expressive gesture synthesis for embodied conversational agents. In Gibet, S., Courty, N., and Kamp, J.-F., editors, *Gesture in Human-Computer Interaction and Simulation*, pages 188–199, Berlin, Heidelberg. Springer Berlin Heidelberg.

[95] Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243–259.

[96] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). β-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations*.

[97] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

[98] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv e-prints*, page arXiv:1207.0580.

[99] Hinton, G. E. and Zemel, R. S. (1993). Autoencoders, minimum description length and Helmholtz free energy. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, page 3–10, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[100] Hortensius, R., Hekele, F., and Cross, E. S. (2018). The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):852–864.

[101] Hudlicka, E. and Gunes, H. (2012). Benefits and limitations of continuous representations of emotions in affective computing: Introduction to the special issue. *International Journal of Synthetic Emotions*, 3(1):i–vi.

[102] Huis In 't Veld, E. M. J., van Boxtel, G. J. M., and de Gelder, B. (2014a). The body action coding system I: Muscle activations during the perception and expression of emotion. *Social Neuroscience*, 9(3):249–264.

[103] Huis In 't Veld, E. M. J., van Boxtel, G. J. M., and de Gelder, B. (2014b). The body action coding system II: Muscle activations during the perception and expression of emotion. *Frontiers in Behavioral Neuroscience*, 8:330.

[104] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

[105] Häring, M., Bee, N., and Andre, E. (2011). Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots. In *IEEE International Workshop on Robot and Human Interactive Communication*.

[106] Höök, K. (2009). Affective loop experiences: Designing for interactional embodiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3585–3595.

[107] Irfan, B., Lyubova, N., Garcia Ortiz, M., and Belpaeme, T. (2018). Multi-modal open-set person identification in HRI. In *International Conference on Human-Robot Interaction Social Robots in the Wild workshop*.

[108] Itoh, K., Miwa, H., Matsumoto, M., Zecca, M., Takanobu, H., Roccella, S., Carrozza, M. C., Dario, P., and Takanishi, A. (2004). Various emotional expressions with emotion expression humanoid robot WE-4RII. In *IEEE Conference on Robotics and Automation, 2004. TExCRA Technical Exhibition Based.*, pages 35–36.

[109] Jerram, M., Lee, A., Negreira, A., and Gansler, D. (2014). The neural correlates of the dominance dimension of emotion. *Psychiatry Research: Neuroimaging*, 221(2):135 – 141.

[110] Jimenez Rezende, D., Mohamed, S., and Wierstra, D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv e-prints*, page arXiv:1401.4082.

[111] John, S. and Pavel., M. (1974). *Messages of the body*. Free Press New York.

[112] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). *An introduction to variational methods for graphical models*, page 105–161. MIT Press, Cambridge, MA, USA.

[113] Joyce, J. M. (2011). *Kullback-Leibler divergence*, pages 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg.

[114] Kahn, P., Kanda, T., Ishiguro, H., Freier, N., Severson, R., Gill, B., Ruckert, J., and Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology*, 48:303–14.

[115] Kanamori, M., Suzuki, M., and Tanaka, M. (2002). Maintenance and improvement of quality of life among elderly patients using a pet-type robot. *Nihon Ronen Igakkai zasshi. Japanese journal of geriatrics*, 39(2):214. In Japanese.

[116] Karg, M., Samadani, A., Gorbet, R., Kühnlenz, K., Hoey, J., and Kulić, D. (2013). Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing*, 4(4):341–359.

[117] Kassambara, A. (2019). *rstatix: Pipe-friendly framework for basic statistical tests*. R package version 0.3.0. https://CRAN.R-project.org/package=rstatix.

[118] Killgore, W. D. S. and Yurgelun-Todd, D. A. (2007). The right-hemisphere and valence hypotheses: could they both be right (and sometimes left)? *Social Cognitive and Affective Neuroscience*, 2(3):240–250.

[119] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv e-prints*, page arXiv:1412.6980.

[120] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv e-prints*, page arXiv:1312.6114.

[121] Kingma, D. P. and Welling, M. (2019). An Introduction to Variational Autoencoders. *arXiv e-prints*, page arXiv:1906.02691.

[122] Kleinsmith, A. and Bianchi-Berthouze, N. (2007). Recognizing affective dimensions from body posture. In Paiva, A., Prada, R., and Picard, R. W., editors, *Affective Computing and Intelligent Interaction*, pages 48–58. Springer Berlin Heidelberg.

[123] Kleinsmith, A. and Bianchi-Berthouze, N. (2013). Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33.

[124] Kleinsmith, A., Rebai, I., Berthouze, N., and Martin, J.-C. (2009). Postural expressions of emotion in a motion captured database and in a humanoid robot. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, AFFINE '09, New York, NY, USA. Association for Computing Machinery.

[125] Knight, H. and Simmons, R. (2014). Expressive motion with x, y and theta: Laban Effort Features for mobile robots. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 267–273.

[126] Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

[127] Koppensteiner, M. (2011). Perceiving personality in simple motion cues. *Journal of Research in Personality*, 45(4):358–363.

[128] Kraemer, H. C. (1981). Extension of Feldt's approach to testing homogeneity of coefficients of reliability. *Psychometrika*, 46(1):41–45.

[129] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

[130] Kühnlenz, K., Sosnowski, S., and Buss, M. (2010). Impact of animal-like features on emotion expression of robot head EDDIE. *Advanced Robotics*, 24(8-9):1239–1255.

[131] Laban, R. (1964). *Modern educational dance*. Macdonald & Evans Ltd.

[132] Lakatos, G., Gácsi, M., Konok, V., Brúder, I., Bereczky, B., Korondi, P., and Miklósi, Á. (2014). Emotion attribution to a non-humanoid robot in different social situations. *PLoS ONE*, 9.

[133] Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford University Press USA.

[134] Le Cun, Y., Jackel, L. D., Boser, B., Denker, J. S., Graf, H. P., Guyon, I., Henderson, D., Howard, R. E., and Hubbard, W. (1989). Handwritten digit recognition: Applications of neural network chips and automatic learning. *IEEE Communications Magazine*, 27(11):41–46.

[135] LeDoux, J. (2012). Rethinking the emotional brain. *Neuron*, 73(4):653–676. Publisher: Elsevier.

[136] Lee, K., Peng, W., Jin, S., and Yan, C. (2006). Can robots manifest personality? An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *Journal of Communication*, 56:754–772.

[137] Leite, I., Castellano, G., Pereira, A., Martinho, C., and Paiva, A. (2012). Long-term interactions with empathic robots: Evaluating perceived support in children. In Ge, S. S., Khatib, O., Cabibihan, J.-J., Simmons, R., and Williams, M.-A., editors, *Social Robotics*, pages 298–307, Berlin, Heidelberg. Springer Berlin Heidelberg.

[138] Lenth, R. (2019). *emmeans: Estimated marginal means, aka Least-Squares Means*. R package version 1.4.3.01. https://CRAN.R-project.org/package=emmeans.

[139] Li, J. and Chignell, M. (2011). Communication of emotion in social robots through simple head and arm movements. *International Journal of Social Robotics*, 3(2):125–142.

[140] Libin, A. V. and Libin, E. V. (2004). Person-robot interactions from the robopsychologists' point of view: The robotic psychology and robotherapy approach. *Proceedings of the IEEE*, 92(11):1789–1803.

[141] Liddell, T. M. and Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79:328 – 348.

[142] Lim, A. and Okuno, H. G. (2014). A recipe for empathy. *International Journal of Social Robotics*, 7(1):35–49.

[143] Lin, J., Spraragen, M., and Zyda, M. (2012). Computational models of emotion and cognition. In *Advances in Cognitive Systems*, pages 59–76.

[144] Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., and Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*, 35(3):121–143.

[145] Lisetti, C. and Hudlicka, E. (2015). Why and how to build emotion-based agent architectures. In Calvo, R., D'Mello, S., Gratch, J., and Kappas, A., editors, *The Oxford Handbook of Affective Computing*, chapter 8, pages 94–109. Oxford University Press, Inc., USA, 1st edition.

[146] Lucas, J., Tucker, G., Grosse, R., and Norouzi, M. (2019). Don't blame the ELBO! A Linear VAE Perspective on Posterior Collapse. *arXiv e-prints*, page arXiv:1911.02469.

[147] MacLennan, B. J. (2009). Robots react, but can they feel? In Vallverdú, J. and Casacuberta, D., editors, *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*, chapter 8, pages 133–153. IGI Global.

[148] Makhzani, A. and Frey, B. (2013). k-Sparse autoencoders. *arXiv e-prints*, page arXiv:1312.5663.

[149] Mancini, M. and Pelachaud, C. (2008). Distinctiveness in multimodal behaviors. In *Proceedings of the 7th International Joint conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 159–166.

[150] Maréchal, C., Mikolajewski, D., tyburek, K., Prokopowicz, P., Bougueroua, L., Ancourt, C., and Wegrzyn-Wolska, K. (2019). Survey on AI-Based Multimodal Methods for Emotion Detection. In *High-Performance Modelling and Simulation for Big Data Applications*, pages pp 307–324. Springer.

[151] Marmpena, M., Garcia, F., and Lim, A. (2020). Generating robotic emotional body language of targeted valence and arousal with conditional variational autoencoders. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 357–359, New York, NY, USA. Association for Computing Machinery.

[152] Marmpena, M., Lim, A., and Dahl, T. S. (2018). How does the robot feel? Perception of valence and arousal in emotional body language. *Paladyn*, 9(1):168–182.

[153] Marmpena, M., Lim, A., Dahl, T. S., and Hemion, N. (2019). Generating robotic emotional body language with variational autoencoders. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 545–551.

[154] Marsella, S. C., Johnson, W. L., and LaBore, C. (2000). Interactive pedagogical drama. In *Proceedings of the fourth international conference on Autonomous agents*, pages 301–308.

[155] Masuda, M. and Kato, S. (2010). Motion rendering system for emotion expression of human form robots based on laban movement analysis. In Avizzano, C. A. and Ruffaldi, E., editors, *19th IEEE International Conference on Robot and Human Interactive Communication, Viareggio, Italy, RO-MAN, 2010, September 13-15, 2010*, pages 324–329. IEEE.

[156] Masuda, M., Kato, S., and Itoh, H. (2010a). A Laban-based approach to emotional motion rendering for human-robot interaction. In Yang, H. S., Malaka, R., Hoshino, J., and Han, J., editors, *Entertainment Computing - ICEC 2010, 9th International Conference, ICEC 2010, Seoul, Korea, September 8-11, 2010. Proceedings*, volume 6243 of *Lecture Notes in Computer Science*, pages 372–380. Springer.

[157] Masuda, M., Kato, S., and Itoh, H. (2010b). Laban-based motion rendering for emotional expression of human form robots. In Kang, B.-H. and Richards, D., editors, *Knowledge Management and Acquisition for Smart Systems and Services*, pages 49–60, Berlin, Heidelberg. Springer Berlin Heidelberg.

[158] Matsui, D., Minato, T., MacDorman, K. F., and Ishiguro, H. (2018). Generating natural motion in an android by mapping human motion. In Ishiguro, H. and Dalla Libera, F., editors, *Geminoid Studies: Science and Technologies for Humanlike Teleoperated Androids*, pages 57–73. Springer Singapore, Singapore.

[159] McColl, D. and Nejat, G. (2014). Recognizing emotional body language displayed by a human-like social robot. *International Journal of Social Robotics*, 6(2):261–280.

[160] McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., and Kaliouby, R. e. (2016). AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, page 3723–3726, New York, NY, USA. Association for Computing Machinery.

[161] Mcgraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(4):390-390.

[162] McKinney, W. et al. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

[163] Mehrabian, A. (2007). *Nonverbal communication*. Aldine Transaction, New Brunswick, NJ.

[164] Mehrabian, A. and Russell, J. A. (1974). *An approach to environmental psychology*. The MIT Press. Pages: xii, 266.

[165] Meijer, M. d. (1989). The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, 13(4):247–268.

[166] Miklósi, Á. and Gácsi, M. (2012). On the utilization of social animals as a model for social robotics. *Frontiers in psychology*, 3:75.

[167] Moerland, T. M., Broekens, J., and Jonker, C. M. (2018). Emotion in reinforcement learning agents and robots: A survey. *Machine Learning*, 107(2):443–480.

[168] Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. (2019). Monte Carlo gradient estimation in machine learning. *arXiv e-prints*, page arXiv:1906.10652.

[169] Monceaux, J., Becker, J., Boudier, C., and Mazel, A. (2009). Demonstration: First steps in emotional expression of the humanoid robot Nao. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, pages 235–236, New York, NY, USA. ACM.

[170] Montecillo Puente, F. J. (2010). *Human motion transfer on humanoid robot.* Theses, Institut National Polytechnique de Toulouse - INPT.

[171] Mortillaro, M., Meuleman, B., and Scherer, K. R. (2012). Advocating a componential appraisal model to guide emotion recognition. *International Journal of Synthetic Emotions*, 3(1):18–32.

[172] Moshkina, L. and Arkin, R. C. (2005). Human perspective on affective robotic behavior: a longitudinal study. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1444–1451.

[173] Nakagawa, K., Shinozawa, K., Ishiguro, H., Akimoto, T., and Hagita, N. (2009). Motion modification method to control affective nuances for robots. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5003–5008.

[174] Nakata, T., Mori, T., and Sato, T. (2001). Quantitative analysis of impression of robot bodily expression based on Laban movement theory. *Journal of the Robotics Society of Japan*, 19(2):252–259. In Japanese.

[175] Nomura, T. and Nakao, A. (2010). Comparison on identification of affective body motions by robots between elder people and university students: A case study in Japan. *International Journal of Social Robotics*, 2(2):147–157.

[176] Novikova, J. and Watts, L. (2014). A design model of emotional body expressions in non-humanoid robots. In *Proceedings of the Second International Conference on Human-agent Interaction*, HAI '14, pages 353–360, New York, NY, USA. ACM.

[177] Ochs, M., Niewadomski, R., and Pelachaud, C. (2015). Facial expressions of emotions for virtual characters. In Calvo, R., D'Mello, S., Gratch, J., and Kappas, A., editors, *The Oxford Handbook of Affective Computing*, chapter 18, pages 261–272. Oxford University Press, Inc., USA, 1st edition.

[178] Oliphant, T. E. (2006). *A guide to NumPy*, volume 1. Trelgol Publishing USA.

[179] Ortony, A., Clore, G., and Collins, A. (1988). *The cognitive structure of emotion*. Cambridge University Press.

[180] Paiva, A., Leite, I., and Ribeiro, T. (2015). Emotion modeling for social robots. In Calvo, R., D'Mello, S., Gratch, J., and Kappas, A., editors, *The Oxford Handbook of Affective Computing*, chapter 21, pages 296–300. Oxford University Press, Inc., USA, 1st edition.

[181] Panksepp, J. and Watt, D. (2011). What is basic about basic emotions? Lasting lessons from affective neuroscience. *Emotion Review*, 3(4):387–396.

[182] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

[183] Pelachaud, C. (2005). Multimodal expressive embodied conversational agents. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 683–689.

[184] Picard, R. W. (1997). *Affective Computing*. MIT Press, Cambridge, MA, USA.

[185] Prete, G., Laeng, B., Fabri, M., Foschi, N., and Tommasi, L. (2015). Right hemisphere or valence hypothesis, or both? The processing of hybrid faces in the intact and callosotomized brain. *Neuropsychologia*, 68:94 – 106.

[186] R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

[187] Reisenzein, R. (2015). A short history of psychological perspectives on emotion. In Calvo, R., D'Mello, S., Gratch, J., and Kappas, A., editors, *The Oxford Handbook of Affective Computing*, chapter 2, pages 21–37. Oxford University Press, Inc., USA, 1st edition.

[188] Ribeiro, T. and Paiva, A. (2012). The illusion of robotic life: principles and practices of animation for robots. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 383–390.

[189] Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011). Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 833–840, USA. Omnipress.

[190] Rossi, S., Cimmino, T., Matarese, M., and Raiano, M. (2019). Coherent and incoherent robot emotional behavior for humorous and engaging recommendations. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–6.

[191] Rossi, S., Staffa, M., and Tamburro, A. (2018). Socially assistive robot for providing recommendations: Comparing a humanoid robot with a mobile application. *International Journal of Social Robotics*, 10:265–278.

[192] Rumbell, T., Barnden, J., Denham, S., and Wennekers, T. (2012). Emotions in autonomous agents: Comparative analysis of mechanisms and functions. *Autonomous Agents and Multi-Agent Systems*, 25(1):1–45.

[193] Ruocco, M., Larafa, M., and Rossi, S. (2019). Emotional distraction for children anxiety reduction during vaccination. *arXiv e-prints*, page arXiv:1909.04961.

[194] Russell, J. A. (1993). Forced-choice response format in the study of facial expression. *Motivation and Emotion*, 17(1):41–51.

[195] Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145.

[196] Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76 5:805–19.

[197] Saad, E., Broekens, J., Neerincx, M. A., and Hindriks, K. V. (2019). Enthusiastic robots make better contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1094–1100.

[198] Saldien, J., Goris, K., Vanderborght, B., Vanderfaeillie, J., and Lefeber, D. (2010). Expressing emotions with the social robot Probo. *International Journal of Social Robotics*, 2(4):377–389.

[199] Scherer, K. (2009). Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364:3459 – 3474.

[200] Scherer, K., Schorr, A., and Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

[201] Schulz, T., Torresen, J., and Herstad, J. (2019). Animation techniques in human-robot interaction user studies: A systematic literature review. *ACM Transactions on Human-Robot Interaction*, 8(2):12:1–12:22.

[202] Sharkey, A. and Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and information technology*, 14(1):27–40.

[203] Sharma, M., Hildebrandt, D., Newman, G., Young, J. E., and Eskicioglu, R. (2013). Communicating affect via flight path: Exploring use of the Laban Effort System for designing affective locomotion paths. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 293–300.

[204] Shibata, T., Mitsui, T., Wada, K., Touda, A., Kumasaka, T., Tagami, K., and Tanie, K. (2001). Mental commit robot and its application to therapy of children. In *2001 IEEE/ASME International Conference on Advanced Intelligent Mechatronics.*, volume 2, pages 1053–1058 vol.2.

[205] Shoemake, K. (1985). Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '85, pages 245–254, New York, NY, USA. ACM.

[206] Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

[207] Silva, R., Louro, L., Malheiro, T., Erlhagen, W., and Bicho, E. (2016). Combining intention and emotional state inference in a dynamic neural field architecture for human-robot joint action. *Adaptive Behavior*, 24(5):350–372.

[208] Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 3483–3491. Curran Associates, Inc.

[209] Sparrow, R. and Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, 16(2):141–161.

[210] Suguitan, M., Bretan, M., and Hoffman, G. (2019). Affective robot movement generation using CycleGANs. In *14th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2019, Daegu, South Korea, March 11-14, 2019*, pages 534–535.

[211] Takahashi, K., Hosokawa, M., and Hashimoto, M. (2010). Remarks on designing of emotional movement for simple communication robot. In *2010 IEEE International Conference on Industrial Technology*, pages 585–590.

[212] Tamura, T., Yonemitsu, S., Itoh, A., Oikawa, D., Kawakami, A., Higashi, Y., Fujimooto, T., and Nakajima, K. (2004). Is an entertainment robot useful in the care of elderly people with severe dementia? *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(1):M83–M85.

[213] Thimmesch-Gill, Z., Harder, K. A., and Koutstaal, W. (2017). Perceiving emotions in robot body language: Acute stress heightens sensitivity to negativity while attenuating sensitivity to arousal. *Computers in Human Behavior*, 76:59 – 67.

[214] Tsiourti, C., Weiss, A., Wac, K., and Vincze, M. (2017). Designing emotionally expressive robots: A comparative study on the perception of communication modalities. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, page 213–222, New York, NY, USA. Association for Computing Machinery.

[215] Turkle, S. (2010). In good company? On the threshold of robotic companions. In *Close Engagements with Artificial Companions*, pages 3–10. John Benjamins.

[216] van den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel recurrent neural networks. *arXiv e-prints*, page arXiv:1601.06759.

[217] Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S*. Springer, New York, fourth edition.

[218] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA. ACM.

[219] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R.,

Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . . (2020). Scipy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272.

[220] Wada, K., Shibata, T., Saito, T., and Tanie, K. (2004). Effects of robot-assisted activity for elderly people and nurses at a day service center. *Proceedings of the IEEE*, 92(11):1780–1788.

[221] Walk, R. and Walters, K. L. (1988). Perception of the smile and other emotions of the body and face at different distances. *Bulletin of the Psychonomic Society*, 26(6):510–510.

[222] Wallbott, H. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, 28:879–896.

[223] Wallbott, H. G. and Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of personality and social psychology*, 51(4):690.

[224] Wasala, K., Gomez, R., Donovan, J., and Chamorro-Koc, M. (2019). Emotion specific body movements: Studying humans to augment robots' bodily expressions. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, OZCHI'19, page 503–507, New York, NY, USA. Association for Computing Machinery.

[225] Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070.

[226] White, T. (2016). Sampling generative networks. *arXiv e-prints*, page arXiv:1609.04468.

[227] Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

[228] Wickham, H., François, R., Henry, L., and Müller, K. (2019). *dplyr: A grammar of data manipulation*. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr.

[229] Witkower, Z. and Tracy, J. L. (2019). Bodily communication of emotion: Evidence for extrafacial behavioral expressions and available coding systems. *Emotion Review*, 11(2):184–193.

[230] Xu, J., Broekens, J., Hindriks, K., and Neerincx, M. A. (2015). Mood contagion of robot body language in human robot interaction. *Autonomous Agents and Multi-Agent Systems*, 29(6):1216–1248.

[231] Yohanan, S. and MacLean, K. E. (2011). Design and assessment of the Haptic Creature's affect display. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 473–480.

[232] Yoon, Y., Ko, W.-R., Jang, M., Lee, J., Kim, J., and Lee, G. (2019). Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proceedings of The International Conference in Robotics and Automation (ICRA)*.

[233] Zhao, L. and Badler, N. I. (2001). Synthesis and acquisition of Laban movement analysis qualitative parameters for communicative gestures. *Technical Reports (CIS)*, page 116.

[234] Zhou, S., Zelikman, E., Lu, F., Ng, A. Y., Carlsson, G., and Ermon, S. (2020). Evaluating the Disentanglement of Deep Generative Models through Manifold Topology. *arXiv e-prints*, page arXiv:2006.03680.

Appendix A

Hand-designed animations

In this Appendix, we provide supplementary material for the study we presented in Chapter 4. More specifically, in Table A.1 we list the 36 hand-coded animations and their properties, and in Table A.2 we list these animations with their valence and arousal labels.

Table A.1 Hand-designed animations: properties and informal descriptions. Note: Tag = animation category assigned by the animators, Duration in seconds, Speed = total frames divided by key frames, Modality is coded as M = Motion, L = LEDs, S = Sound.

Tag	Description	Duration	Speed	Modality
Happy_4	Slightly raises both hands, elbows bent, head nods up and down	2.4	6.67	ML
Interested_2	Bends right elbow and brings hand near waist, while left hand raises near mouth. Head raises looking upward and on the left	2.56	2.34	MLS
Joy_1	Both hands in front of the torso, elbows bent, then open wide, head looks upward	2.8	3.93	M
Loving_01	Both hands brought behind the body, torso bends to right and then to the left	4.36	6.65	ML
Confident_1	Bends both elbows slightly while nodding head up and down	4.96	3.83	ML

Continued on next page

Table A.1 – continued from previous page

Tag	Description	Duration	Speed	Modality
Heat_1	Looks downward, hands a bit open, bends torso slightly forward	4.68	1.92	M
Optimistic_1	Slightly nodding and then looks upward to the left	6.6	2.58	ML
Peaceful_1	Bends both elbows and raises hands slightly up and then down. Head moves downward and then up respectively.	3.44	2.62	ML
Content_01	Head looks up, torso bends right and left, hands open and close repeatedly	2.44	14.34	ML
Excited_01	Elbows bent, hands raise in front midway and move slightly up and down repeatedly, while head slightly nodding	2.24	16.07	M
Happy_01	Swings left and right, with hands on the sides and head nodding up and down	3.48	6.61	M
Joyful_01	Swings left and right, while moving arms up and down in front	3.56	8.99	M
Angry_3	Turns head to the right and brings left hand further away from body to the left	2.32	4.31	MS
Frustrated_1	Turns head to the right, nods right and left, hands brought in front, slightly bent	3.08	4.22	MLS
Hot_1	Head nods right and left	1.96	2.04	M
SadReaction_01	Moves torso back, then forward to the right, head nods left and right	2.88	9.38	M

Continued on next page

Table A.1 – continued from previous page

Tag	Description	Duration	Speed	Modality
Angry_4	Moves backward, raises left hand up and shakes it while head is looking up	2.68	7.84	MS
Fear_1	Arms raise slightly in front, head looks up, then right and left	4.4	5.23	MLS
Fearful_1	Both hands raise in front of the head, left hand higher	6.52	3.68	ML
Sad_01	Hands raise slightly in front with elbows bent, shakes head right and left and downward on the right direction	3.88	7.47	ML
Bored_01	Right hand is brought in front of the mouth and slightly moves back and forth	3.68	4.08	ML
Disappointed_1	Torso moves forward and slightly downward, while head nods right and left	2.88	3.82	ML
Lonely_1	A slight movement of hands and head up and then downward. Hands on the sides	7.04	2.98	ML
Shocked_1	Hands and head move slightly up and then downward. Hands in front	4.2	3.33	ML
AskForAttention_3	Brings right hand in front of the mouth, and then down while looking right	4.24	2.59	MLS
Chill_01	Slightly nodding and swinging left and right	6.72	3.57	M
Puzzled_1	Brings right hand in front of the mouth and left hand on waist. Head to the left	4	2.75	ML

Continued on next page

Table A.1 – continued from previous page

Tag	Description	Duration	Speed	Modality
Relaxation_2	Raises both hands midways in front with elbows bent and looks up	6.56	2.44	ML
Curious_01	Bends torso forward left and then right, hands open on the sides	2.68	2.24	M
SurprisedBig_01	Hands open on the sides, looks up	2.52	9.13	ML
Surprised_01	Slightly raises hands and looks up	3.08	3.57	M
Surprised_1	Raises both hands midway in front with elbows bent, looks around	4.88	4.92	M
Alienated_1	Torso slightly bent forward, head up, arms hanging	10.36	1.93	M
Hesitation_1	Repeatedly looks up to the left and down to the right, while right hand raises and falls again	9.04	2.32	MLS
Innocent_1	Hands meet in front of the torso and head looks left and right upward	7.68	2.47	ML
Stretch_2	Hands open on the sides, torso bends backward and head looks up	7.28	2.61	M

Table A.2 The final affect labels. Mean across 20 raters for valence and arousal, with the averaged confidence rates in judgment (C_Mean), as well as the original categorical tags of the animations and the pre-assigned classes of valence/arousal levels combinations (Neg = Negative, Neu = Neutral, Pos = Positive, Tir= Tired, Cal = Calm, Exc = Excited).Valence and arousal ratings range from 0 to 1, with 100 points resolution, while confidence rates where collected with 5-point Likert scales.

Animation		Arousal		Valence	
Tag	Class	Mean(SD)	C_Mean(SD)	Mean(SD)	C_Mean(SD)
Happy_4	Pos/Cal	0.69(0.15)	3.95(0.76)	0.73(0.19)	3.85(1.14)
Interested_2	Pos/Cal	0.79(0.10)	4.15(0.49)	0.78(0.18)	4.10(0.79)

Continued on next page

Table A.2 – continued from previous page

Animation		Arousal		Valence	
Tag	Class	Mean(SD)	C_Mean(SD)	Mean(SD)	C_Mean(SD)
Joy_1	Pos/Cal	0.55(0.20)	3.50(1.15)	0.65(0.15)	3.35(1.18)
Loving_01	Pos/Cal	0.63(0.26)	3.45(1.23)	0.57(0.22)	3.00(1.08)
Confident_1	Pos/Tir	0.67(0.18)	3.55(0.76)	0.70(0.15)	3.60(0.75)
Heat_1	Pos/Tir	0.33(0.22)	3.60(0.60)	0.41(0.14)	3.65(0.81)
Optimistic_1	Pos/Tir	0.49(0.24)	3.70(0.66)	0.62(0.16)	3.30(1.08)
Peaceful_1	Pos/Tir	0.40(0.25)	3.45(1.32)	0.51(0.12)	2.90(1.07)
Content_01	Pos/Exc	0.85(0.14)	4.05(0.94)	0.71(0.23)	3.45(1.10)
Exc_01	Pos/Exc	0.81(0.16)	4.15(0.67)	0.71(0.27)	3.70(1.22)
Happy_01	Pos/Exc	0.82(0.12)	4.30(0.73)	0.86(0.12)	4.65(0.49)
Joyful_01	Pos/Exc	0.80(0.13)	3.85(0.99)	0.64(0.25)	3.75(0.97)
Angry_3	Neg/Cal	0.63(0.23)	3.90(0.55)	0.40(0.20)	3.40(0.88)
Frustrated_1	Neg/Cal	0.61(0.23)	3.25(0.97)	0.20(0.17)	4.00(0.79)
Hot_1	Neg/Cal	0.35(0.20)	3.45(1.00)	0.32(0.18)	3.75(0.72)
SadReaction_01	Neg/Cal	0.32(0.29)	3.90(0.97)	0.19(0.15)	4.05(0.83)
Angry_4	Neg/Exc	0.78(0.25)	4.05(1.00)	0.19(0.22)	3.65(1.42)
Fear_1	Neg/Exc	0.89(0.17)	4.15(0.75)	0.43(0.32)	3.50(1.10)
Fearful_1	Neg/Exc	0.90(0.11)	4.40(0.60)	0.14(0.16)	4.20(0.89)
Sad_01	Neg/Exc	0.75(0.17)	3.80(0.83)	0.33(0.22)	3.55(0.89)
Bored_01	Neg/Tir	0.48(0.31)	3.65(1.09)	0.49(0.13)	2.90(1.12)
Disappointed_1	Neg/Tir	0.37(0.19)	3.40(0.94)	0.21(0.14)	4.10(0.72)
Lonely_1	Neg/Tir	0.28(0.23)	3.65(0.81)	0.28(0.19)	3.70(1.17)
Shocked_1	Neg/Tir	0.48(0.28)	3.70(0.80)	0.56(0.17)	2.85(0.93)
AskForAttention_3	Neu/Cal	0.49(0.25)	3.45(0.89)	0.38(0.22)	3.45(0.94)
Chill_01	Neu/Cal	0.56(0.18)	3.65(0.88)	0.64(0.18)	3.70(0.80)
Puzzled_1	Neu/Cal	0.60(0.19)	3.80(0.62)	0.49(0.12)	3.20(0.83)
Relaxation_2	Neu/Cal	0.50(0.22)	3.15(1.14)	0.56(0.15)	2.60(1.27)

Continued on next page

Table A.2 – continued from previous page

Animation		Arousal		Valence	
Tag	Class	Mean(SD)	C_Mean(SD)	Mean(SD)	C_Mean(SD)
Curious_01	Neu/Exc	0.89(0.13)	4.25(0.79)	0.62(0.17)	3.50(1.15)
SurprisedBig_01	Neu/Exc	0.82(0.16)	4.00(0.73)	0.40(0.19)	3.15(0.93)
Surprised_01	Neu/Exc	0.78(0.19)	4.00(0.79)	0.46(0.22)	3.20(1.11)
Surprised_1	Neu/Exc	0.71(0.15)	3.45(0.94))	0.60(0.15)	3.25(1.02)
Alienated_1	Neu/Tir	0.20(0.24)	4.05(0.60))	0.41(0.16)	3.05(1.05)
Hesitation_1	Neu/Tir	0.70(0.16)	3.55(1.00))	0.32(0.22)	3.15(1.09)
Innocent_1	Neu/Tir	0.52(0.21)	4.05(0.39))	0.69(0.15)	3.50(0.89)
Stretch_2	Neu/Tir	0.40(0.21)	3.40(1.05))	0.60(0.18)	3.10(0.97)

Appendix B

Experimental interface

In this Appendix, we present three views from the participant's interface used in the experiment described in Chapter 7.

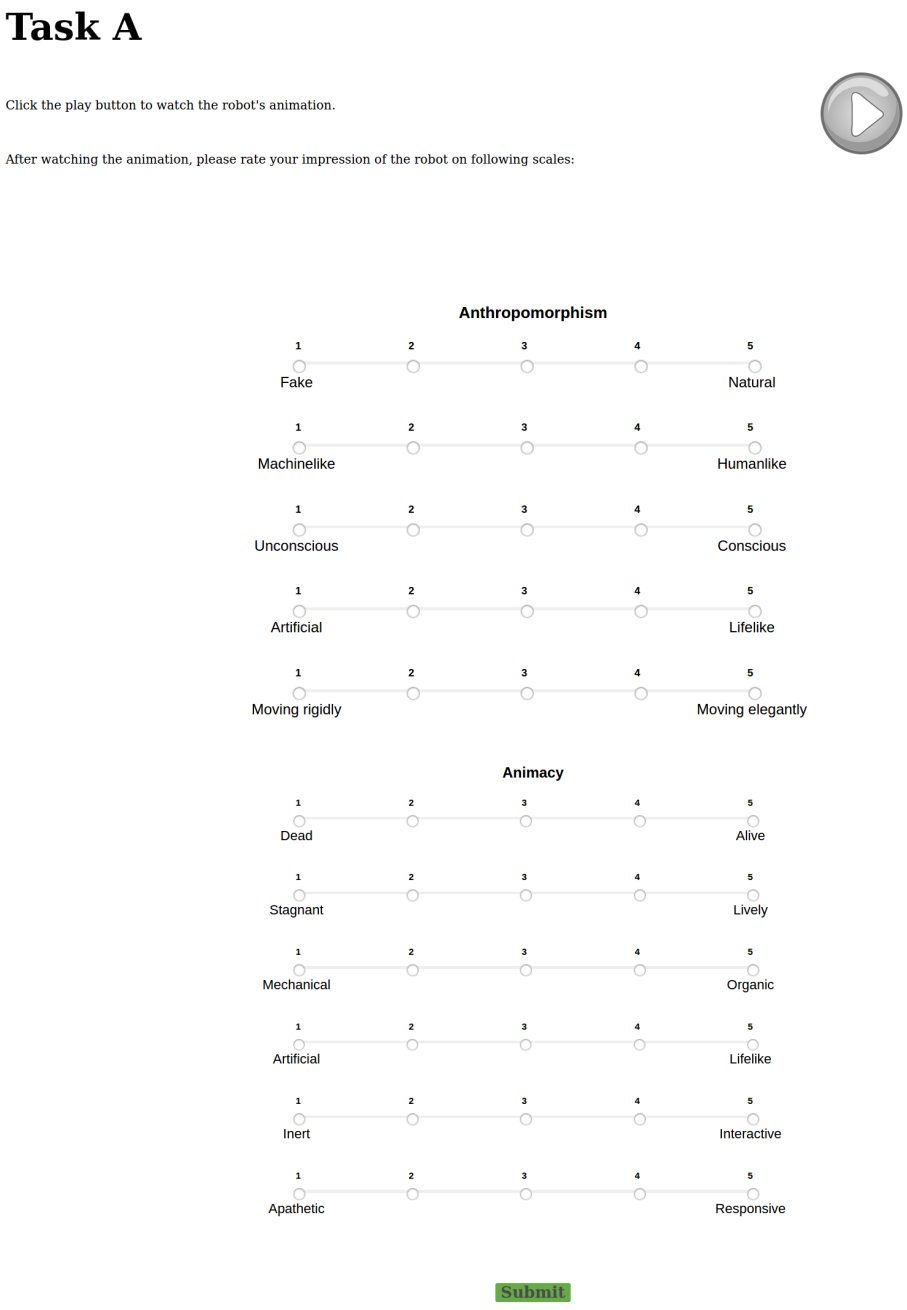

Fig. B.1 The web interface used for Part A and C of the experimental session described in Chapter 7. It contains a play button with which the participant activates the robot to display a set of concatenated animations. Subsequently, the participant submits her scores on two scales of Godspeed Questionnaire [16], Anthropomorphism and Animacy.

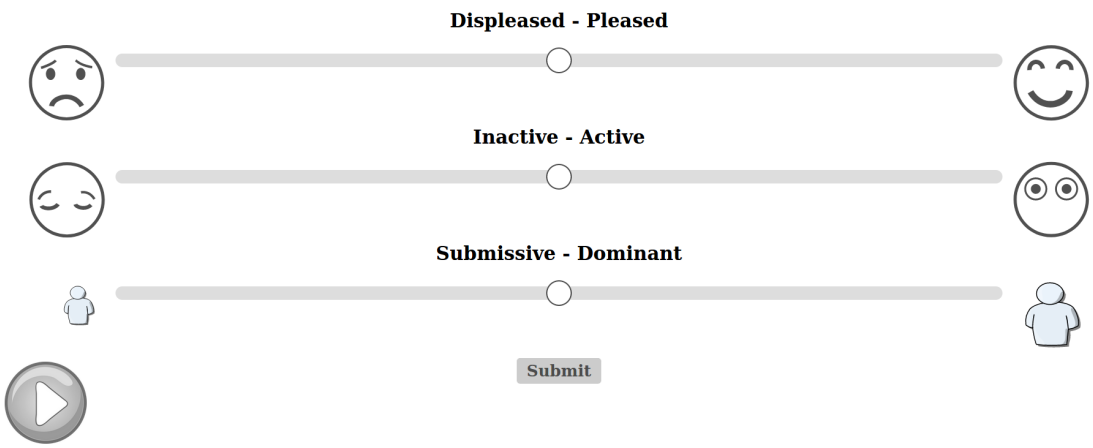

Fig. B.2 The first view of Part B of the experimental session described in Chapter 7. This play button activates the robot to display a single generated animation. Then, the participant submits her valence, arousal and dominance scores on the sliders. The range of each slider is from 0 to 1 with 100 steps.

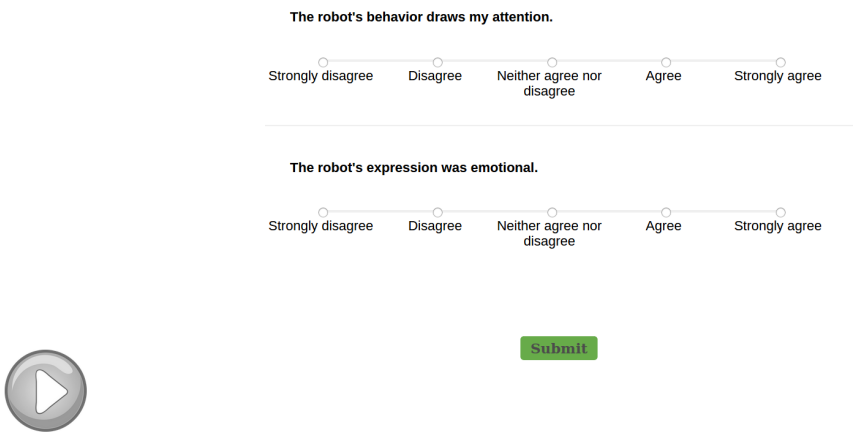

Fig. B.3 The second view of Part B of the experimental session described in Chapter 7. The play button allows the user to replay the animation she evaluated in the first view. Then, the participant submits her scores on the Likert scales.